Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Una aproximación a la predicción del valor de acciones en la bolsa de Valores aplicando técnicas de Data Mining.

Trabajo Fin de Máster

Máster Universitario en Software y Sistemas

Autor: Javier Isaac Espinosa Muñoz

Director: Francisco Javier Segovia Pérez

Madrid, Julio de 2015

Departamento de Lenguajes y Sistemas Informáticos e Ingeniería de Software (DLSSIIS)
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo S/N
28660 – Boadilla del Monte (Madrid)
Correo electrónico: master.muss@fi.upm.es

Predicción del valor de acciones en la bolsa Mexicana de Valores aplicando técnicas de Data Mining.

Trabajo de Fin de Máster
Máster Universitario en Software y Sistemas

Autor: **Javier Isaac Espinosa Muñoz**
Ingeniero en Mecatrónica

Director: **Francisco Javier Segovia Pérez**
Doctor en Informática

Madrid, Julio de 2015

“Make things as simple as possible, but not simpler.”

Albert Einstein (1879 – 1955)

“Los pájaros nacidos en jaula creen que volar es una enfermedad”

Alejandro Jodorowsky (1929 – Actualidad)

“My favorite things in life don't cost any money. It's really clear that the most precious resource we all have is time”

Steve Jobs (1955 – 2011)

AGRADECIMIENTOS

A mí familia por otorgarme la confianza de seguir trabajando y sobre todo el apoyo de mi director de trabajo de fin de máster el Dr. Francisco Javier Soriano Camino.

RESUMEN

La predicción del valor de las acciones en la bolsa de valores ha sido un tema importante en el campo de inversiones, que por varios años ha atraído tanto a académicos como a inversionistas. Esto supone que la información disponible en el pasado de la compañía que cotiza en bolsa tiene alguna implicación en el futuro del valor de la misma. Este trabajo está enfocado en ayudar a un persona u organismo que decida invertir en la bolsa de valores a través de gestión de compra o venta de acciones de una compañía a tomar decisiones respecto al tiempo de comprar o vender basado en el conocimiento obtenido de los valores históricos de las acciones de una compañía en la bolsa de valores. Esta decisión será inferida a partir de un modelo de regresión múltiple que es una de las técnicas de *datamining*. Para llevar conseguir esto se emplea una metodología conocida como *CRISP-DM* aplicada a los datos históricos de la compañía con mayor valor actual del NASDAQ.

ABSTRACT

The prediction of the value of shares in the stock market has been a major issue in the field of investments, which for several years has attracted both academics and investors. This means that the information available in the company last traded have any involvement in the future of the value of it. This work is focused on helping an investor decides to invest in the stock market through management buy or sell shares of a company to make decisions with respect to time to buy or sell based on the knowledge gained from the historic values of the shares of a company in the stock market. This decision will be inferred from a multiple regression model which is one of the techniques of data mining. To get this out a methodology known as CRISP-DM applied to historical data of the company with the highest current value of NASDAQ is used.

KEYWORDS Data Mining, Stock Investing, Multiple Regression, CRISP-DM.

Índice general

Índice de figuras

Índice de ecuaciones

Índice de tablas

1 Capítulo: Introducción y objetivos

1.1 Planteamiento del problema

La predicción del valor de las acciones en la bolsa de valores se ha convertido en un problema clásico al que se enfrentan académicos e inversionistas, esto es debido a la complejidad que presenta ésta incógnita debido al enorme número de condiciones y variables involucradas en todo el ecosistema que rodea este proceso de evaluación. Uno de los factores de reto al que se enfrenta durante el transcurso de la investigación en este trabajo es la propiedad que presenta esta variable de parecer ser aleatoria a corto plazo, esto debido a su naturaleza no lineal, complicando así la tarea de encontrar un modelo que sea lo suficiente preciso como para confiar en el (Wang, 2003) . Debido a este fenómeno es común que los inversionistas recurran a la estrategia convencional de comprar la acción de su preferencia, mantenerla sin realizar movimientos durante un determinado periodo de tiempo y luego buscar recuperar su inversión y su ganancias mediante la venta de la misma *buy and hold strategy* , algunas veces el resultado de esta esto puede generar un posible retorno negativo que involucran la pérdida del capital inicial del inversor(Greemblatt, 2005), por esta razón los inversores han tratado de encontrar la mejor manera de predecir los cambios en el valor de las acciones de manera que puedan saber cuándo es el mejor momento para vender o comprar. Para conseguir este objetivo, algunos académicos utilizan técnicas de análisis fundamental (Seng & Hancock, 2012),donde las reglas de compraventa *trading* están desarrolladas en la información asociada a las variables macroeconómicas, industriales y de la propia compañía. Otros académicos (Seng & Hancock, 2012) proponen que el análisis fundamental asume que el precio de la acción depende de su valor real *enterprise value* y de la expectativa de retorno de inversión *return on investment (ROI)*.

Algunos otros trabajos de investigación emplean técnicas de análisis técnico (Murphy, 1999), en el cual las reglas de *trading* se desarrollan en base a los datos históricos del precio de la acción durante un periodo de tiempo. Con esto se

intenta predecir el valor futuro de los movimientos usando la información de los valores en el pasado. Este acercamiento está basado en la suposición que la historia se repite en sí misma y que las futuras direcciones del mercado pueden determinarse a partir de examinar estos datos. Así, se asume que las tendencias y patrones de precio sobre las acciones se pueden identificar, por lo que se puede obtener un beneficio de ello, si esta presunción es correcta de que el análisis técnico puede desembocar en encontrar patrones existentes en los precios, es entonces en principio usar técnicas de *data mining* para descubrir estos patrones y poder predecir los futuros precios de las acciones en la bolsa.

1.2 Objetivos

Como se ha planteado en el apartado anterior, el propósito del presente trabajo es generar un modelo que ayude al inversor mediante reglas útiles a encontrar una estrategia óptima basado en su preferencia a la hora de apostar por una compañía, reduciendo las pérdidas de capital y maximizando las ganancias durante un periodo determinado. Una vez realizado este modelo se evaluara usando perfiles de inversión, para finalmente comparar el rendimiento obtenido frente al método tradicional de comprar y mantener o *buy and hold*(Chong, Ling, Ng, Yat, & Muhamad, 2014) hasta concretar su venta al final tiempo de inversión. Por lo que el gran reto es usar los datos disponibles en la bolsa y encontrar patrones escondidos y generar los modelos más adecuados, de manera tradicional el método para encontrar este valor en la información era generado a través de un análisis manual, pero esto se ha vuelto impráctico por que el volumen de datos existente crece de manera exponencial, este trabajo utiliza una de las técnicas para generar predicciones utilizando *data mining,* la cual es regresión múltiple esto para hacer uso de los datos históricos y poder encontrar relación entre las compañías que estén involucradas con la compañía a la que se le quiera invertir, con esto poder ayudar al inversor al momento de tomar la decisión de comprar más acciones o de venderlas con la finalidad de obtener ganancias durante un plazo de inversión estipulado.

1.3 Hipótesis

Considerando lo expuesto hasta este punto, el presente trabajo de fin de máster tiene como objetivo principal verificar la siguiente hipótesis:

"Realizando un análisis técnico de los datos históricos del valor de las acciones durante un plazo determinado, es posible predecir el cambio de precio de una acción, con la finalidad de ayudar al inversor al momento de tomar la decisión de compraventa *trading* de sus acciones, utilizando para esto regresión múltiple"

Para la verificación de la misma, además de comprobar su viabilidad de implementación se propondrá un perfil de estrategia que reduzca las perdidas y aumente las ganancias utilizando el modelo obtenido y comparándolo con la estrategia tradicional de comprar y mantener sin realizar algún movimiento durante un periodo determinado de inversión, evaluando los resultados obtenidos en cada uno de los casos.

1.4 Estructura y limitación del trabajo de fin de máster

Analizar datos históricos del valor de las acciones en una bolsa de valores puede significar miles o hasta millones de instancias, por lo cual los datos utilizados en este trabajo son únicamente de las compañías catalogas como *big cap*[1] y *large cap*[2] del NASDAQ, otra de las posibles limitaciones del trabajo es que en las operaciones de compraventa *trading* no se considera la comisión que se realiza con cada movimiento de la acción ya que este valor varía en función de cada empresa que toma el rol de intermediario entre el inversionista y la casa de bolsa también conocido como *market maker*[3] que en el caso del NASDAQ son aproximadamente 500 compañías[4].

Seguido a este punto se procede a detallar la estructura del presente trabajo de fin de máster, la cual se organiza en 5 capítulos. El capítulo 2 llamado *Estado de la cuestión* es enfocado en analizar la literatura referente al tema acerca de

[1] Capitalización de mercado mayor a 200 billones de dólares
[2] Capitalización de mercado mayor a 10 billones de dólares
[3] http://www.nasdaqtrader.com/Trader.aspx?id=MarketMakerProcess
[4] http://www.nasdaqomx.com/transactions/markets/commodities/markets/market-makers

usar técnicas de *data mining* con la intención de predecir su valor. El capítulo 3 es referente a la metodología usada para conseguir el modelo de predicción. El capítulo 4 desarrolla la etapa de experimentación donde los datos recolectados pasaran por el modelo de manera que puedan ser evaluados. El capítulo 5 cuenta de manera crítica los resultados obtenidos durante el desarrollo del trabajo. Finalmente, en el capítulo 6 se detalla una breve conclusión, además de proponer mejores para la continuación del trabajo.

2 Capítulo: Estado de la cuestión

En este capítulo se describe de forma general cuales han sido las técnicas y teorías más sobresalientes para el desarrollo de este trabajo de fin de máster. Primero se detalla la historia de la teoría financiera moderna. Posteriormente se describe, algunas técnicas de *data mining* que han sido ampliamente utilizadas en esta área a la par del estado actual.

2.1 Teoría Financiera Moderna

2.1.1 Historia

Uno de los primeros economistas neoclásicos conocido como Irving Fisher (Fisher, 1907) ,estableció bases para realizar finanzas cuantitativas de la última parte del siglo con el propósito de diseñar un curso de la conducta racional y científica para los participantes del mercado de valores. En París, el matemático francés Louis Bachelier estudió los cambios de precios en el mercado de valores y se dio cuenta de que la incertidumbre y la aleatoriedad de los cambios de precios podrían ser manejados con la distribución de Gauss que a principios del siglo 20 ya estaba ganando interés y las aplicaciones de la física en el estudio de infinita cantidad de causas independientes, él consciente de las limitaciones de uso en el modelado de la conducta humana, ya que, se dio cuenta que en un grupo, las acciones humanas ya no son independientes entre sí, sino que son propensos a que su conducta cambie en base a lo que escuchan a su alrededor. En su tesis doctoral (Bachelier, 1906) ,descubrió las matemáticas del movimiento browniano y afirmó que los precios históricos no pueden predecir el futuro, y los cambios de precios se describe mejor como si fuera un andar aleatorio. Casi cincuenta años más tarde, Harry Markowitz ,utilizo la programación lineal, que fue ampliamente utilizada en la investigación de operaciones en el problema de la optimización de portafolio de un inversor descrito en su artículo "*Portfolio Selection*" (Markowitz, 1952).

2.1.2 Teorías de predicción del mercado de acciones

Al momento de predecir los valores de las acciones, se encuentran dos importantes teorías disponibles. La primera Hipótesis del mercado eficiente *efficient market hypothesis (EMH)* propuesta por Eugene Fama (1964) y la segunda es la teoría del andar aleatorio *Random Walk Theory* propuesta por Burton Malkiel (1973).

2.1.2.1 Hipótesis del mercado eficiente (EMH)

La contribución que logro hacer fama con el desarrollo de esta hipótesis fue bastante significativa, el EMH establece que el valor actual de la acción es reflejo de la asimilación de toda la información disponible. Esto significa que, dada la información, no existe posibilidad de realizar una predicción de que el futuro del valor pueda cambiar de alguna manera, de esta manera cuando entra nueva información al sistema el estado de desequilibrio es inmediatamente descubierto por lo que se procede a corregir el valor (B. Malkiel & Fama, 1970).Está teoría se separa en tres vertientes(Schumaker & Chen, 2008):

1) En la forma débil EMH, sólo el precio y la información histórica está embebido en el valor actual. Este tipo de EMH descarta cualquier forma de predicción que utilice como fundamento sólo los datos históricos del valor de la acción, ya que esta variable mantiene un comportamiento aleatorio por lo que los valores futuros tienen una correlación de cercana a cero.
2) En la forma semi-fuerte EMH, en está además del precio y la información histórica se incorpora también lo actual, esto incluye reportes sobre negociaciones adicionales, así como los datos de volumen, los indicadores fundamentales y el pronóstico de ventas del trimestre.
3) En la forma fuerte EMH, incluye la información histórica pública, así privada, cómo lo es reportes internos de rendimiento, en el precio de acción.

2.1.2.2 Teoría del andar aleatorio

Una perspectiva diferente en cuanto a la predicción viene de parte de la Teoría de andar aleatorio *Random Walk Theory* desarrollada por Malkiel en 1973, en esta teoría la predicción del valor de las acciones es imposible ya que supone que los precios están determinados al azar y superar el mercado de valores es visto como inviable. Está teoría tiene una similitud con la base teórica semi-Fuerte de EMH donde se supone que toda la información es publica y está disponible para todos. Sin embargo, está teoría declara qué aún con el conocimiento de esta información, la predicción continua siendo inviable.

2.1.3 Enfoque hacia la predicción del mercado de acciones

A partir de las teorías de EMH y *Random Walk*, han surgido dos filosofías distintas respecto a la manera en que se realiza la compraventa *trading* de acciones ha surgido. Estos dos enfoques son el análisis técnico y el análisis fundamental.

2.1.3.1 Enfoque hacia el análisis técnico

El término denota un acercamiento básico donde los valores históricos son estudiados, utilizando tablas como principal herramienta. Esto permite que se puedan proponer modelos basado en reglas o patrones de los valores históricos que también se le conocer como *financial time-series*. El principio básico incluye conceptos como la naturaleza de tendencia de precios, confirmación y divergencia, así como el efecto de volumen negociado. Desde su comienzo se han desarrollado un gran número de métodos de predicción del valor de las acciones(Hellstrom, 1998), en la actualidad se siguen desarrollando, un claro ejemplo es este mismo trabajo el cual está sustentado en estos principios.

El análisis técnico se basa en datos de series temporales numéricas por lo que intenta pronosticar el mercado de acciones usando indicadores resultantes de aplicar análisis técnico, tiene como fundamento una hipótesis bastante aceptada que dice que todas las reacciones del mercado a todas las noticias están contenidas en los precios en tiempo real de las acciones. Debido a esto el analista técnico ignora las noticias. Su principal preocupación es identificar las tendencias existentes y anticipar las tendencias futuras del mercado de valores

de los gráficos. Pero gráficos o series temporales de datos numéricos sólo contienen el evento y no la causa por qué sucedieron(Kroha & Baeza-Yates, 2004).

En el análisis técnico, se cree que la sincronización del mercado es crítica y las oportunidades pueden encontrarse a través de calcular una media del valor histórico y del movimiento de volumen, para finalmente compararlos en contra de los precios actuales. Los técnicos utilizan gráficos y técnicas de modelado para identificar tendencias en precio y volumen. Se basa en datos históricos para predecir resultados futuros(Schumaker & Chen, 2010).

2.1.3.2 Enfoque hacia el análisis fundamental

El análisis fundamental(Thomsett, 1998), investiga los factores que afectan la oferta y la demanda. El objetivo es reunir e interpretar esta información y actuar antes de que la información se vea reflejada en el precio de las acciones. El lapso de tiempo entre un evento y su respuesta en el mercado, representa una oportunidad comercial. El análisis fundamental se basa en los datos económicos de las empresas y trata de predecir los mercados utilizando los datos económicos que las empresas tienen que publicar con regularidad, es decir, los informes anuales y trimestrales, informes de auditoría, balances, cuentas de resultados, etc. Las noticias tienen también una gran importancia para los inversores, ya que si están usando el análisis fundamental, puede que las noticias contengan factores que pueden afectar la oferta y la demanda con lo que habría un cambio de valor en la acción no esperado.

En la filosofía de comercio fundamentalista, el precio de un valor puede ser determinado a través de los aspectos prácticos de los números financieros. Estos números se derivan de la economía en general, del sector de la industria en particular, o más típicamente, a partir de la propia empresa. Figuras como la inflación, el retorno de inversión *return on investment (ROE)* y los niveles de deuda pueden desempeñar un papel en la determinación del valor de una acción(Schumaker & Chen, 2009).

Una de las áreas de éxito limitado en la predicción del mercado de valores proviene de datos de texto y el uso de artículos de prensa en la predicción de

precios. Información sobre el informe o noticias de última hora de la empresa puede afectar dramáticamente el precio de las acciones. Ha habido muchas investigaciones llevadas a cabo para investigar la influencia de las noticias en el mercado de valores y la reacción de estos hacia los comunicados de prensa. Los estudios generales muestran que la bolsa reacciona a las noticias y los resultados obtenidos a partir de estudios anteriores indican que los artículos de noticias afectan el movimiento del mercado de valores.

2.2 Descubrimiento de conocimiento en base de datos

También conocido como "*knowledge discovery in database*" fue inventado en el primero KDD *workshop* en 1989(Frawley, Piatetsky-shapiro, & Matheus, 1992), con el motivo de enfatizar que el conocimiento es producto de llevar a cabo un descubrimiento en los datos. Descubrimiento de conocimiento *knowledge discovery* está definido como una extracción no trivial de lo implícito, desconocido y potencialmente ser una información útil proveniente de los datos.

Atreves de una amplia variedad de campos, los datos están siendo recogidos y acumulados a un ritmo espectacular. Hay una necesidad urgente de una nueva generación de teorías y herramientas computacionales que nos ayuden a los seres humanos en la extracción de información útil (conocimiento) obtenido del rápido crecimiento de volúmenes de datos digitales. Estas teorías y herramientas están sujetas al crecimiento del campo emergente de descubrimiento de conocimiento en bases de datos. KDD es la intersección de los campos de investigación como lo es aprendizaje automático, reconocimiento de patrones, bases de datos, las estadísticas, inteligencia artificial (IA), la adquisición de conocimientos de los sistemas expertos, la visualización de datos y computación de alto rendimiento(Fayyad, Piatetsky-Shapiro, & Smyth, 1996a).

La minería de datos *data mining* es un paso en el proceso KDD que consiste en la aplicación de análisis de datos y de algoritmos de descubrimiento de manera que, bajo las limitaciones de eficiencia computacional aceptables, producen una enumeración particular de patrones (o modelos) sobre los datos, este es un concepto que se ha consolidado desde finales de 1980. Esto cubre una amplia

gama de técnicas para el descubrimiento eficiente de esta valiosa información, al no ser evidente en grandes colecciones de datos. En esencia, *data mining* tiene que ver con el análisis de datos y el uso de técnicas de software para encontrar patrones y regularidades en los conjuntos de datos(Hand & Hand, 1998).

El proceso de KDD consta de muchos pasos, incluyendo la selección de datos, procesamiento previo, la transformación, la minería de datos y la evaluación, todos los pasos se pueden repetir en múltiples iteraciones. (Figura 2-1).

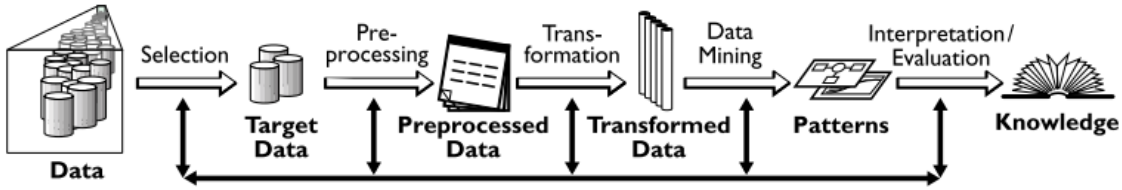

Figura 2-1 Vista general de los pasos que constituyen un proceso de KDD
Fuente:(Fayyad, Piatetsky-Shapiro, & Smyth, 1996b)

El término también se ha usado en un contexto negativo en el cual se agota los datos mediante pruebas de una multitud de variables sin ningún objetivo a priori ni algún razonamiento hipotético de correlación o causalidad, si no que hasta que alguna combinación se ajusta a los datos(Foster & Kesselman, 2004). Además de estas preocupaciones, el uso de técnicas de DM en descubrir por completo la nueva información y las dimensiones de los datos debería ser una práctica completamente aceptable en la ciencia. Otra vertiente hace hincapié en la capacidad de construir modelos econométricos de forma automatizada de acuerdo a un algoritmo de reglas de decisión. Miles de regresiones y evaluaciones modelo pueden realizarse en cuestión de segundos, la inferencia estadística puede ser automatizada de acuerdo con las propiedades de los datos, y las decisiones de política puede hacerse y ajustarse en tiempo real con la llegada de nuevos datos. También hace una observación sobre el reto importante que las investigaciones se enfrentan en la incorporación de pensar y métodos económica en el modelo automatizado y el proceso de selección de variables(Phillips, Haven, & Foundation, 2006).

Data mining métodos se pueden subdividir en dos clases distintas, predictivos y descriptivos, que también se llaman supervisadas y no supervisadas, respectivamente Figura 2-2. En los métodos sin supervisión ninguna variable de destino se identifica para el algoritmo de DM pero los patrones y estructuras entre todas las variables se buscan. Estos se utilizan a menudo en la reducción de la dimensión de los datos como la agrupación y el análisis de componentes principales *principal components analysis (PCA)*. Los datos se describen en nuevas dimensiones. Por otra parte en los métodos supervisados se encuentra en particular una variable de destino, y el algoritmo se entrena con los datos para ajustar los parámetros del modelo para las mejores propiedades predictivas. El objetivo es para predecir el nivel del objetivo o clasificar a alguna categoría predefinida. Los más importantes son los árboles de decisión y redes neuronales, sino también a los modelos clásicos, como regresión logística y lineal.

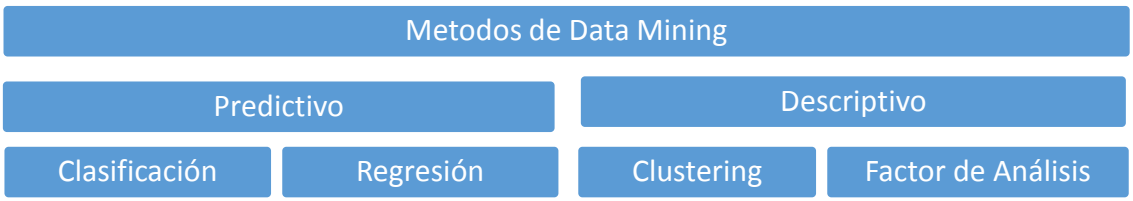

Figura 2-2 Métodos de *Data Mining*

2.2.1 Data Mining aplicado a la bolsa de valores

En la última década han sucedido algunos cambios importantes en el entorno de los mercados financieros. El desarrollo de poderosas instalaciones de comunicación y de comercio ha ampliado el alcance de selección para los inversores. Predecir el retorno de inversión de una acción es un tema financiero importante que ha atraído la atención de los investigadores durante muchos años. Se trata de una suposición de que la información histórica de la empresa catalogada como fundamental se encuentra a disposición del público de manera que tengan una oportunidad de realizar modelos o estrategias de manera que les permita realizar una estimación de los futuros rendimientos de las acciones (Enke & Thawornwong, 2005).

Por esa razón, varios investigadores se han centrado en el análisis técnico y el uso avanzado de matemáticas y ciencias. Amplia atención se ha dedicado al

campo de la inteligencia artificial, así como a desarrollar técnicas de *data mining* (WANG & CHAN, 2006). Algunos modelos se han propuesto e implementado usando las técnicas antes mencionadas Algunos modelos(Tsang et al., 2007), realizaron un estudio empírico armando un sistema de alerta compraventa *trading* utilizando redes neuronales de propagación *back propagation neuronal network* (BPNN), su NN fue el nombre en código NN5. El sistema fue entrenado y probado con datos de acciones pasadas de Hong Kong hacia el banco de Shanghai durante el período comprendido entre enero de 2004 y diciembre de 2005. Los resultados empíricos muestran que el sistema implementado fue capaz de predecir direcciones de movimiento de precios a corto plazo con exactitud de 74 %. Con esto podemos examinar la eficacia de los modelos de redes neuronales utilizados por el nivel de estimación y clasificación. El resultado muestra que las estrategias comerciales orientadas por los modelos de clasificación usando una red neural generan mayores ganancias bajo la misma exposición al riesgo de que los sugeridos por otras estrategias.

2.2.2 CRISP-DM

Existe una multitud de diferentes especificaciones para seguir un proceso de *data mining* (Olson & Delen, 2008).Sin embargo todos comparten el punto de inicio en un dispositivo electrónico de datos, por ejemplo una base de datos. El usar una metodología de *data mining* nos permite estar seguros de que el esfuerzo estará conducido en todo momento hacia un modelo establece y con capacidad de mejora y de réplica por otra persona o institución(Han & Kamber, 2006),esto se logra organizando el proceso de recopilación de datos, análisis de datos, difusión de resultados, implementando resultados y hacer el seguimiento del resultado con el objetivo de proponer mejoras. Para construir el modelo de análisis de la bolsa de valores usando la técnica de regresión múltiple usaremos CRISP *(Cross- Industry Standard Process for data mining)*(Ibm, 2010).Está metodología fue propuesta a mediados de los 90's por un consorcio de compañías europeas para servir como un modelo para realizar un proceso estándar no-propietario en el área de *data mining*. El ciclo de vida de un proyecto de *data mining* usando está metodología hace uso de seis fases mostrado a continuación (Figura 2-1). Se puede notar que la secuencia no es rígida, moverse

entre las etapas entre fases es casi siempre requerido. La salida de cada fase o de una tarea en particular determina que es lo que toca hacer después. Las flechas indican la frecuencia más común entre dependencias.

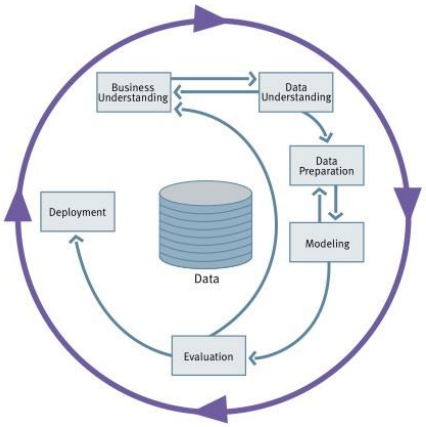

Figura 2-3 Etapas del modelo de referencia CRISP-DM

El circulo de la figura 2 muestra la naturaleza del *data mining*, en este no termina una vez que se realiza una implementación, si no que mediante las lecciones aprendidas durante el proceso y de la solución implementada pueden desencadenar en nuevas preguntas en la fase de negocio que no se cubrieron la primera vez, así que el proceso se beneficia de las experiencias anteriores. A continuación, se describe brevemente cada fase:

Comprensión del negocio "*Business understanding"*

Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, a continuación, convertir este conocimiento en una definición del problema de *data mining*, por consiguiente poder desarrollar un plan preliminar para lograr estos objetivos.

Comprensión de los datos "*Data understanding"*

Esta fase involucra realizar una comprensión junto con la primera adquisición de datos, prosigue con las actividades que permiten familiarizarse con la información, identificar problemas de calidad de datos, detectar subconjuntos interesantes para formar hipótesis sobre la información oculta en estos datos iniciales.

Preparación de los datos “*Data preparation”*

Esta fase cubre todas las actividades necesarias para construir el conjunto de datos finales [este conjunto de dato es el que se introduce en la herramienta de modelado], esto siempre de los datos iniciales. Esta tarea de preparación de datos es susceptible de ser realizada múltiples veces y puede ser en cualquier orden prescrito. Las tareas incluyen seleccionar las tablas, registro, selección de atributos, así como la transformación y la limpieza de datos para que se encuentre listo para usarse en la herramienta de modelado.

Modelado “*Modeling”*

En esta fase, se seleccionan varias técnicas de modelado, así como se aplican sobre los datos provenientes de la fase anterior, los parámetros de los modelos pasan a estar calibrados con sus valores óptimos. Típicamente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos en la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

Evaluación “*Evaluation”*

En esta etapa del proyecto, ya se han construido un modelo o varios que parecen tener alta calidad desde una perspectiva de análisis de datos. Antes de proceder a la implementación del modelo final, es importante evaluar a fondo y revisar los pasos ejecutados para crearlo, esto para estar seguro de que el modelo logra adecuadamente los objetivos de negocio. Un objetivo clave es determinar si hay algún problema en cuanto al alcance y cumplimiento de los objetivos de negocio importantes y que no se encuentren considerados hasta este momento. Al final de esta fase, se debe alcanzar una decisión sobre el uso de estos resultados que provienen de nuestro modelo de *data mining*.

Implementación *“Deployment”*

En esta etapa final generalmente la creación del modelo no es el final del proyecto. Aunque el propósito del modelo sea incrementar el conocimiento de los datos, este conocimiento obtenido necesita presentarse de manera organizada, de una manera que el cliente puede utilizarla los conocimientos

adquiridos. A menudo implica la utilización de modelos considerados vivos *live,* ya que se utilizan en la toma de decisión de una organización como por ejemplo, la personalización en tiempo real de anuncios de algunas páginas Web con las bases de datos de marketing(Perlich et al., 2012). Dependiendo de los requisitos, la fase de despliegue puede ser tan simple como la generación de un informe o tan complejo como la implementación de un proceso de *data mining* en toda la empresa. En muchos casos, es el cliente, no el analista de datos, él que lleva a cabo los pasos de implementación. Sin embargo, si el analista el responsable de llevar a cabo el esfuerzo de despliegue, es importante dar a conocer al cliente los detalles y ajustes necesarios que necesitan ser llevado a cabo con el fin de hacer realidad el uso de los modelos creados.

3 Capítulo: Materiales y Métodos

3.1 Introducción

Este capítulo implementa el uso de las primeras tres fases de la metodología CRISP-DM: comprensión del negocio, comprensión de los datos y preparación de los datos.

3.2 Fase I: *Comprensión del negocio*.

La razón y objetivo principal del presenta trabajo de fin de máster consiste en la definición de un modelo general a partir de los valores históricos en el mercado de valores. Se han identificado una serie de pasos comunes deben realizarse independientemente de la acción que se esté tratando en cada momento.

3.2.1 Objetivos del negocio

- Minimizar las pérdidas del portafolio del inversor en lo posible.
- Maximizar las ganancias del portafolio del inversor en lo posible.

3.2.2 Criterios de éxito del negocio

- Mejorar el desempeño en términos de ganancias del portafolio del inversor, respecto al uso de mantener la estrategia de comprar y mantener *buy and hold strategy*.

3.2.3 Inventario de recursos

La herramienta de *data mining* usada en este trabajo es STATGRAPHICS Centurion [5] . Esta solución está enfocada al análisis de datos y visualización en la plataforma Microsoft Windows[6] , con el uso de este instrumento podremos desarrollar los modelos predictivos necesarios y la implementación en los entornos operativos de nuestro cliente que en este

[5] http://www.statgraphics.com/centurion-xvii

[6] https://www.microsoft.com/en-us/windows

caso es el inversionista con el objetivo de ayudar a mejorar la toma de decisiones.

3.2.4 Objetivos de *data mining*

- Predecir los cambios en el valor de las acciones durante el periodo de inversión elegido.

3.2.5 Criterios de éxito de *data mining*

- Maximizar el margen de exactitud de las predicciones de la estimación del valor de la acción.

3.2.6 Plan de Proyecto

A continuación se presenta una proyección de las fases con sus respectivas tareas, marcando la duración requerida para completar cada paso. Tabla 3-1.

Tabla 3-1 Plan de proyecto

Tarea	**Duración en horas**
1.Comprensión de los datos	
1.1. Reporte de descripción de datos	8
1.2. Reporte de calidad de datos	4
2.Preparación de los datos	
2.1. Inclusión / Exclusión de datos	8
2.2. Reporte de limpieza de datos	20
2.3. Derivación de atributos	20
2.4. Unificación de datos	20
2.5. Formateo de los datos	12
3. Modelado	
3.1. Técnica de modelado	8
3.2. Diseño de la prueba	20
3.3. Modelo	8
3.4. Evaluar el modelo	12
4. Evaluación	
4.1. Evaluar el resultado de *data minig* respecto a los criterios de éxito del negocio	4
4.2. Revisión del proceso	12

4.3. Lista de posibles acciones	4
5. Implementación	
5.1. Plan de implementación	20
5.2. Plan de monitorización y mantenimiento	16
Total de horas invertidas	196

3.3 Fase II: *Comprensión de los datos*

Aunque todas las fases del modelo propuesto son importantes, quizás esta fase de comprensión de los datos de las fuentes de información constituye un elemento fundamental. Al fin y al cabo, los datos obtenidos en esta fase son los que se usarán para extraer la información que interese en cada momento. Una selección errónea de los datos conllevará a un irremediable fracaso incluso aunque se dispongan de los mejores métodos de análisis. Si la información relevante o que interesa no está contenida en las fuentes seleccionadas, resultará imposible obtenerla a partir de las mismas. Esta última afirmación puede resultar obvia pero conviene tenerla siempre presente al iniciar cualquier proceso de obtención de información.

3.3.1 Colección inicial de datos

El conjunto de datos utilizados para esta fase, están contenidos en la base de datos del *National Association of Securities Dealers Automated Quotations (NASDAQ)* contiene el precio histórico de 3,063[7] compañías enlistadas en esta bolsa desde el 8 de febrero de 1971 día que comenzó con las operaciones de bolsa. Esta lista de datos disponibles es bastante larga y variable, la razón de esto es que nuevas empresas se agregan cada año, así como también dejan de cotizar en la misma. La figura siguiente muestra la clasificación por industria y su valor en el mercado. Tabla 3-2.

[7] Información valida al martes 30 de junio de 2015; fuente : http://www.nasdaq.com/screening/companies-by-region.aspx?region=ALL&exchange=NASDAQ

Tabla 3-2 Clasificación de las compañias en el NASDAQ

Industria	N° compañias	Valor en USD	Porcentaje
n/a	282	$ 152,224,931,576.01	2%
Basic Industries	78	$ 55,563,817,509.27	1%
Capital Goods	183	$ 215,499,205,446.41	2%
Consumer Durables	83	$ 75,447,540,058.21	1%
Consumer Non-Dur.	104	$ 256,796,404,981.66	3%
Consumer Services	369	$ 1,881,683,727,473.55	22%
Energy	80	$ 61,624,004,753.95	1%
Finance	603	$ 517,776,818,993.46	6%
Healthcare	593	$ 1,329,473,781,823.79	15%
Miscellaneous	99	$ 245,080,675,126.22	3%
Public Utilities	68	$ 80,291,551,810.66	1%
Technology	465	$ 3,644,142,785,568.86	42%
Transportation	56	$ 110,369,203,845.52	1%
Total	3063	$8,625,974,448,967.58	100%

La forma de elegir a las compañías fue que pertenecieran a una misma industria por lo que se eligió la que tuviera mayor porcentaje de valor en la bosa, que en este caso es tecnología *technology* en la cuales están clasificadas 465 compañías, este número de empresas continua siendo bastante grande, por lo que utilizaremos el criterio de elegir aquellas compañías consideradas como como grandes *big cap* y largas *large cap*. Tabla 3-3.

Tabla 3-3 Clasificación Grandes y Largas Compañias de Tecnología del NASDAQ

Simbolo	Compañia	Accion	Market Cap USD
AAPL	Apple Inc.	122.57	$ 706,129,447,100.00
GOOGL	Google Inc.	541.7	$ 370,169,237,827.00
MSFT	Microsoft Corporation	44.24	$ 357,882,810,254.48
GOOG	Google Inc.	516.83	$ 353,174,390,227.30
FB	Facebook, Inc.	85.65	$ 240,525,434,469.90
INTC	Intel Corporation	29.5	$ 139,948,000,000.00
CSCO	Cisco Systems, Inc.	26.99	$ 137,268,136,822.70
QCOM	QUALCOMM Incorporated	61.91	$ 100,886,624,033.47
TXN	Texas Instruments Incorporated	49.53	$ 51,529,198,062.81
BIDU	Baidu, Inc.	184.58	$ 51,140,212,862.20
ADBE	Adobe Systems Incorporated	79.99	$ 39,806,653,466.26

ADP	Automatic Data Processing, Inc.	80.03	$	37,576,008,120.90
CTSH	Cognizant Technology Solutions Corp.	59.03	$	36,038,853,514.79
YHOO	Yahoo! Inc.	37.23	$	34,938,220,529.64
AVGO	Avago Technologies Limited	128.05	$	33,258,388,085.00
BRCM	Broadcom Corporation	51.2	$	30,617,600,000.00
INTU	Intuit Inc.	103.43	$	28,512,429,982.94
CERN	Cerner Corporation	67.805	$	23,330,230,216.38
AMAT	Applied Materials, Inc.	18.55	$	22,846,148,056.90
NXPI	NXP Semiconductors N.V.	92.235	$	21,537,425,910.00
EA	Electronic Arts Inc.	69	$	21,534,401,958.00
FISV	Fiserv, Inc.	83.41	$	19,750,530,036.15
ADI	Analog Devices, Inc.	61.94	$	19,420,771,968.90
MU	Micron Technology, Inc.	17.63	$	19,067,554,924.84
SWKS	Skyworks Solutions, Inc.	95.6	$	18,269,008,378.40
WDC	Western Digital Corporation	78.31	$	18,082,723,418.60
ATVI	Activision Blizzard, Inc	24.8	$	18,009,420,959.20
SYMC	Symantec Corporation	22.82	$	15,534,211,408.24
ALTR	Altera Corporation	51.2	$	15,415,755,161.60
STX	Seagate Technology.	46.4	$	14,726,503,038.40
CHKP	Check Point Software Technologies Ltd.	78.7	$	14,464,348,001.10
CA	CA Inc.	29.83	$	13,115,580,332.11
LRCX	Lam Research Corporation	77.95	$	12,341,461,188.40
TRIP	TripAdvisor, Inc.	85.58	$	12,281,542,838.84
ADSK	Autodesk, Inc.	52.475	$	11,946,050,926.68
VRSK	Verisk Analytics, Inc.	71.99	$	11,477,459,790.93
SNDK	SanDisk Corporation	54.145	$	11,257,785,029.86

El valor de en conjunto de estas 37 compañías es de $3,083,810,558,902.91 USD lo que representa un 85% del valor de las compañías que están clasificadas en tecnología o un 36% del total del valor de mercado del NASDAQ.De estás 37 acciones seleccionaremos nuestro objetivo, qué en este caso será por el criterio tamaño del mercado. Esta compañía es Apple Inc. (*AAPL)* está en la subcategoría de manufactura de sistemas computacionales. El periodo seleccionado será del 29-Junio-2009 al 30-Junio-2015 un periodo aproximado a seis años de los cuales corresponden 1,512 días de actividades.

3.3.2 Descripción de datos

Los indicadores que fungen como atributos se han mantenido a lo largo de este tiempo tienen las características mostradas a continuación. Tabla 3-4.

Tabla 3-4 Descripción de los atributos

Atributo	Descripción
Fecha	Nos indica el momento en el que fueron tomado los datos, este valor puede ser el año, mes, semana, día, hora y por minuto.
Apertura	Esto indica el precio apertura del día por la acción
Alto	Esto indica el precio máximo pagado durante el periodo por la acción
Bajo	Esto indica el precio mínimo pagado durante el periodo por la acción
Cierre	Esto indica el último precio pagado durante el periodo por la acción
Volumen	Esto indica el número de acciones comercializadas de la acción durante el periodo
Ajuste de cierre	Esto indica el precio ajustado incluyendo cualquier distribución o acción corporativa que se produjera en cualquier momento antes de abrir al periodo siguiente.

3.3.3 Calidad de datos

Los datos obtenidos deberán de contener al menos la mínima información requerida para afrontar el problema, en caso contrario se buscara otra solución de manera que se mantenga uniformidad y coherencia en los datos.

Para esto, investigaremos las características de los datos de cada compañía para comprobar la integridad de los valores. Tabla 3-4.

Tabla 3-5 Verificación de datos por compañia

Simbolo	Compañia	N° instancias	N° errores
AAPL	Apple Inc.	1512	0
GOOGL	Google Inc.	1512	0
MSFT	Microsoft Corporation	1512	0
GOOG	Google Inc.	318	0
FB	Facebook, Inc.	783	0
INTC	Intel Corporation	1512	0
CSCO	Cisco Systems, Inc.	1512	0

QCOM	QUALCOMM Incorporated	1512	0
TXN	Texas Instruments Incorporated	1512	0
BIDU	Baidu, Inc.	1512	0
ADBE	Adobe Systems Incorporated	1512	0
ADP	Automatic Data Processing, Inc.	1512	0
CTSH	Cognizant Technology Solutions Corp.	1512	0
YHOO	Yahoo! Inc.	1512	0
AVGO	Avago Technologies Limited	1485	0
BRCM	Broadcom Corporation	1512	0
INTU	Intuit Inc.	1512	0
CERN	Cerner Corporation	1512	0
AMAT	Applied Materials, Inc.	1512	0
NXPI	NXP Semiconductors N.V.	1233	0
EA	Electronic Arts Inc.	1512	0
FISV	Fiserv, Inc.	1512	0
ADI	Analog Devices, Inc.	1512	0
MU	Micron Technology, Inc.	1512	0
SWKS	Skyworks Solutions, Inc.	1512	0
WDC	Western Digital Corporation	1512	0
ATVI	Activision Blizzard, Inc	1512	0
SYMC	Symantec Corporation	1512	0
ALTR	Altera Corporation	1512	0
STX	Seagate Technology.	1512	0
CHKP	Check Point Software Technologies Ltd.	1512	0
CA	CA Inc.	1512	0
LRCX	Lam Research Corporation	1512	0
TRIP	TripAdvisor, Inc.	895	0
ADSK	Autodesk, Inc.	1512	0
VRSK	Verisk Analytics, Inc.	1442	0
SNDK	SanDisk Corporation	1512	0

De las 37 compañías, las siguientes 6 no tiene la misma cantidad de instancias:

Tabla 3-6 Verificación de calidad en los datos

Símbolo	**Compañía**	**N° instancias**
GOOG	Google Inc.	318
FB	Facebook, Inc.	783
AVGO	Avago Technologies Limited	1485
NXPI	NXP Semiconductors N.V.	1233
TRIP	Trip Advisor, Inc.	895
VRSK	Verisk Analytics, Inc.	1442

3.4 Fase III: *Preparación de los datos*

A pesar de que la preparación de datos es una sola fase en la metodología CRISP-DM, es la que en general consume la mayoría del esfuerzo durante el desarrollo del proceso. El objetivo es unir, transformar, manipular los datos de manera que podamos obtener datos de alta calidad para nuestra siguiente fase de *data mining*.

3.4.1 Inclusión / Exclusión de datos

Los atributos obtenidos de cada compañía son los siguientes:

- Fecha
- Apertura
- Alto
- Bajo
- Cierre
- Volumen
- Ajuste de Cierre

Seleccionaremos sólo el atributo de ajuste de cierre, la razón es que incluye un ajuste en caso de cualquier movimiento antes del siguiente día que vuelva abrir el mercado(Alraddadi, 2015).

3.4.2 Limpieza de datos

Durante la tarea de revisar la calidad de los datos, nos dimos cuenta de que algunas compañías presentaban datos incompletos debido a que no tenía los suficientes días de cotización a la fecha del estudio, normalmente cuando no se encuentra un 10% de datos en comparación a las demás variables es posible ignorar excepto cuando está perdida de datos no es aleatoria(Kock & Verville, 2012) y al no tener posibilidad de estimar estos datos, tomaremos la decisión de eliminar las siguientes 6 compañías:

- GOOG (Google Inc.)
- FB (Facebook, Inc.)
- AVGO (Avago Technologies Limited)

- NXPI (NXP Semiconductors N.V.)
- TRIP (TripAdvisor, Inc.)
- VRSK (Verisk Analytics, Inc)

Por consecuencia nos quedaremos con los datos de 31 compañías de las 37 iniciales.

3.4.3 Derivación de atributos

Al ser un problema de predicción, vamos a proponer usar los valores históricos de una semana (5 días hábiles de transacción), con la intención de convertirlos en atributos, hacer eso necesitaremos convertir un único atributo en 6 atributos, para esto usaremos los datos que provienen de la tarea de limpieza:

Tabla 3-7 Muestra de datos proveniente de la tarea de limpieza.

Date	AAPL
6/29/2009	19.05101
6/30/2009	19.11274
7/1/2009	19.16642
7/2/2009	18.78934
7/6/2009	18.60013
7/7/2009	18.16938

Nuestro objetivo es estimar el valor de la acción de mañana con base a nuestra hipótesis la cual utiliza los valores anteriores. Ecuación 3-1.

Ecuación 3-1 hipótesis para realizar la predicción del valor de la acción

$$AAPL\ (\ mañana\ ) = \alpha * AAPL\ (hoy) + \beta * AAPL\ (ayer) + \cdots$$
$$AAPL\ (\ t\ ) = \alpha * AAPL\ (t\text{-}1) + \beta * AAPL\ (t-2) + \cdots$$

Entonces la tabla quedaría de la siguiente manera:

Tabla 3-8 Derivación de atributos

Date	**AAPL**	**AAPL_1**	**AAPL_2**	**AAPL_3**	**AAPL_4**	**AAPL_5**
7/7/2009	18.16938	18.60013	18.78934	19.16642	19.11274	19.05101

Este mismo procedimiento se realiza para todos los datos de las 31 compañías.

3.4.4 Unificación de datos

Al realizar la derivación para cada compañía, proseguimos con la unión de los datos de todas las compañías en una misma base de datos, por lo que obtenemos de resultados 156 atributos, incluyendo a nuestro objetivo AAPL.

3.4.5 Formateo de los datos

Cómo ultima tarea es exportar los datos unificados a un archivo de tipo valores separados por coma (CSV) para que pueda ser utilizado por la herramienta de *data mining.*

4 Capítulo: Experimentación y resultados

4.1 Introducción

En el capítulo anterior revisamos la aplicación de las primeras tres fases de la metodología CRISP-DM, comprensión del negocio, comprensión de los datos y preparación de los datos.

En este capítulo abordaremos las tres fases restantes de la metodología, con especial énfasis en fase IV: modelado, además se muestran los resultados experimentales obtenidos de la aplicación del método presentado en el capítulo anterior. El objetivo es probar la validez del método propuesto y verificar la hipótesis de partida planteada en el apartado 1.3 del presenta trabajo de fin de máster.

4.2 Fase IV: *Modelado*

El propósito de esta fase de modelado de la metodología CRISP-DM es seleccionar, construir y probar los modelos que den soporte al objetivo de la operación. Para este trabajo, regresión múltiple es la técnica de *data mining* que ha sido elegida y será discutida en esta sección.

4.2.1 Técnica de modelado

La decisión de seleccionar la técnica de regresión múltiple es con la intención de construir un modelo estadístico que describa el impacto de dos o más factores cuantitativos que los llamaremos X sobre una variable dependiente Y(Brown, 2009).

El modelo para realizar una regresión lineal es el siguiente:

Ecuación 4-1 Modelo de regresión lineal.Fuente(Tables, 2009)

$$Y = b_0 + b_1 X_1 + \cdots + b_n X_n$$

- Y es la variable dependiente
- Los términos X_i representan las variables independientes o explicativas

- Los coeficientes del modelo b_i son calculados por el programa estadístico, de modo que se minimicen los residuos.

4.2.2 Diseño de la prueba

Para desarrollar la prueba, separamos el conjunto de datos en dos, uno será el conjunto de entrenamiento y el otro el de pruebas.

Nuestro conjunto de datos es de un periodo de 6 años, por lo que lo separamos en un conjunto de entrenamiento y en un conjunto de pruebas, a continuación las características de estos dos.

- Conjunto de entrenamiento, del 29/06/2009 al 29/06/2014,1254 instancias
- Conjunto de prueba, del 30/06/2014 al 30/06/2014, 253 instancias

4.2.3 Modelo

Para hacer encajar el modelo de regresión lo mejor posible con los datos de entrenamiento, usaremos una combinación de métodos para ello.

Con el método de mínimos cuadrados ordinarios el modelo usara todas las variables independientes.

- Mínimos cuadrados ordinarios con constante
- Mínimos cuadrados ordinarios sin constante

Con el método de Selección de paso hacia adelante, realiza una regresión por pasos hacia adelante. Comenzando con un modelo que incluye sólo una constante, el procedimiento usa una de las variables a la vez, previendo que estas serán estadísticamente significativas una vez añadidas. Las variables también pueden ser eliminadas en etapas posteriores si ya no son estadísticamente significativas.

- Selección de paso hacia adelante con constante
- Selección de paso hacia adelante sin constante

Con el método de Selección de paso hacia atrás, realizara una regresión por pasos hacia atrás. Comenzando con un modelo que incluye todas las variables, el procedimiento elimina variables de uno a la vez si no son estadísticamente

significativas. Las variables eliminadas también se pueden añadir al modelo en los pasos posteriores si se vuelven estadísticamente significativa.

- Selección de paso hacia atrás con constante
- Selección de paso hacia atrás sin constante

4.2.4 Descripción del modelo

Los resultados de los modelos de cada una de las pruebas, se encuentran en el anexo. El modelo elegido fue el que utiliza el método de selección de paso hacia adelante el cual tiene el mejor balance entre el número de variables involucradas en la ecuación en este caso son siete, además de qué el modelo explica el 99.9715% de la variabilidad de AAPL :

Ecuación 4-2 Modelo elegido para optimizar

$$AAPL = 0.987545 * AAPL\_1 - 0.0391297 * CTSH\_3 - 0.0848414 * INTU\_3 + 0.101999 * INTU\_4 + 0.0708764 * SWKS\_1 - 0.0586335 * SWKS\_5 + 0.0188428 * CHKP\_4$$

Ahora que tenemos el modelo elegido, podremos realizar una optimización por lo que procedemos a adecuar un conjunto de datos de prueba con los atributos de las variables elegidas por el modelo, de un universo de pasa de 155 atributos iniciales pasamos a sólo 7 atributos finales con la siguiente correspondencia:

- AAPL_X:´ Valor de acción de Apple del día de hoy[8]
- CTSH_X: Valor de acción de Cognizant Technology Solutions Corp[9]
- INTU_X: Valor de acción de Intuit Inc[10]
- SWKS_X: Valor de acción de Skyworks Solutions, Inc[11]
- CHKP_X: Valor de acción de Check Point Software Technologies Ltd[12]
- APPL: Valor de acción de Apple del día de mañana
- MODELO: Valor de la ecuación del modelo elegido.

[8] http://www.nasdaq.com/symbol/aapl
[9] http://www.nasdaq.com/symbol/ctsh
[10] http://www.nasdaq.com/symbol/intu
[11] http://www.nasdaq.com/symbol/swks
[12] http://www.nasdaq.com/symbol/chkp

Tabla 4-1 Modelo evaluado con los atributos del conjunto de prueba

Date	AAPL_1	CTSH_3	INTU_3	INTU_4	SWKS_1	SWKS_5	CHKP_4	AAPL	MODELO
6/30/2014	90.39378	49.71	80.05554	79.49302	45.73041	47.54731	66.55	91.3274	90.3463
7/1/2014	91.3274	49.39	79.89764	80.05554	46.62396	46.47504	66.76	91.90722	91.48174
7/2/2014	91.90722	49.3	80.25291	79.89764	47.38845	46.45518	66.15	91.86791	92.05547
7/3/2014	91.86791	48.91	79.47328	80.25291	47.27924	46.02826	66.87	92.40842	92.16516

Continuamos realizando nuestro objetivo de optimizar nuestro modelo, como primer paso calculamos la suma de errores al cuadrado entre el valor del modelo y la variable objetivo (AAPL) usando la siguiente formula(Dubey, 2012).

Ecuación 4-3 formula para el cálculo de la suma de errores al cuadrado

$$SSE = \sum (valor\ AAPL - valor\ AAPL\ Modelo)^2$$

Cómo segundo paso calculamos la correlación, con el objetivo de conocer la relación que tiene dos grupos de datos(Hansen, 1999).

Ecuación 4-4 formula para el cálculo de la correlación

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Donde:

x: Valor de la instancia dentro del grupo de datos 1

y: Valor de la instancia dentro del grupo de datos 2

$\bar{x}$: Valor promedio del grupo de datos 1

$\bar{y}$: Valor promedio del grupo de datos 2

Cómo tercer paso calculamos el mixing(Corporation, 2011), esto el valor de la correlación dividido por el valor encontrado de la suma de errores al cuadrado:

Ecuación 4-5 formula para encontrar el mixing

$$Mixing = \frac{Correl(X,Y)}{SSE}$$

Estos valores nos ayudaran a la hora de buscar la optimización de los coeficientes que tenemos en nuestra ecuación, para eso utilizaremos Solver[13].

- Minimizar el SSE
- Maximizar la correlación
- Maximizar el mixing

Para encontrar esta optimización cambiaremos las constantes de los coeficientes por variables.

Ecuación 4-6 con constantes en sus coeficientes

$$= 0.987545*B2 - 0.0391297*C2 - 0.0848414*D2 + 0.101999*E2 + 0.0708764*F2 - 0.0586335*G2 + 0.0188428*H2$$

Al hacer el cambio podremos utilizar la herramienta solver para buscar optimizar estas constantes, realizaremos la aplicación de tres modelos de optimización descritos anteriormente con el propósito de encontrar los mejores coeficientes para nuestra ecuación de nuestro modelo.

- AAPL_1= L$3
- CTSH_3= M$3
- INTU_3= N$3
- INTU_4= O$3
- SWKS_1= P$3
- SWKS_5= Q$3
- CHKP_4= R$3

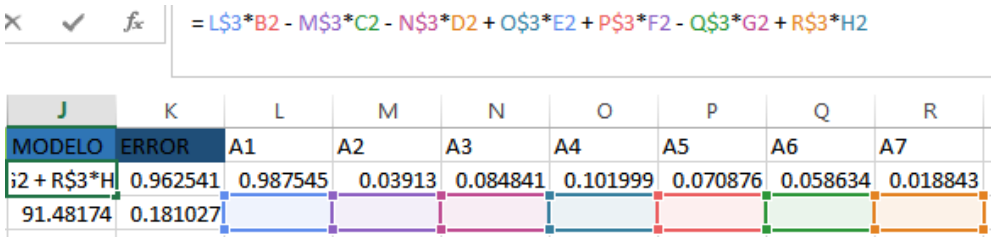

Figure 4-1 Cambio de constantes por variables en los coeficientes de la ecuación del modelo

13 https://support.office.com/en-au/article/Load-the-Solver-Add-in-0e6760e3-dab5-4fd4-bebb-15ee311a4316

Los valores del modelo sin optimizar son los siguientes:

Valores de coeficientes:

A1	A2	A3	A4	A5	A6	A7
0.987545	0.0391297	0.0848414	0.101999	0.070876	0.058634	0.018843

Valores de optimización:

SSE	Correlación	Mixing
638.99492	0.9929858	0.00155398

Grafica del Modelo:

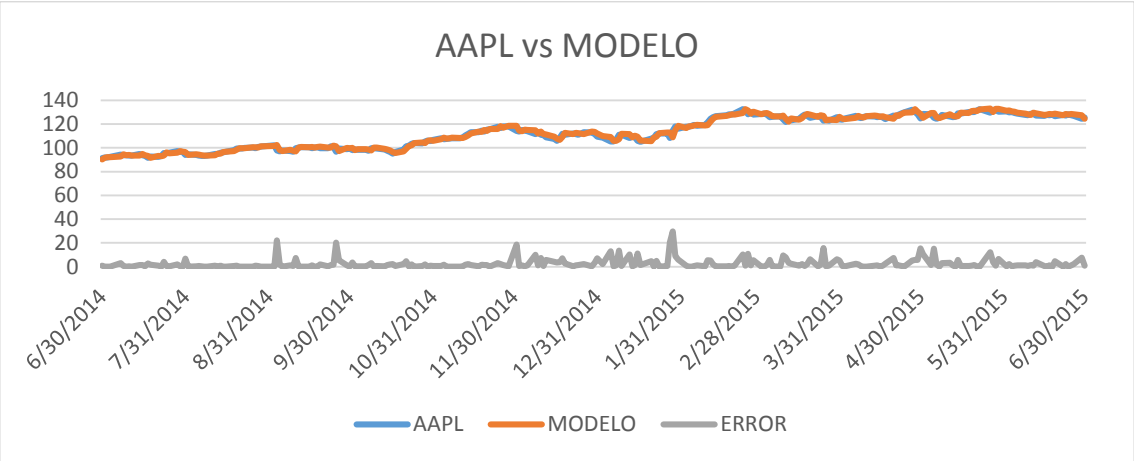

Figure 4-2 Modelo sin Optimizar

4.2.4.1 Ecuación del modelo optimizando SEE

Valores de coeficientes:

A1	A2	A3	A4	A5	A6	A7
0.9717144	-0.0537326	0.02967956	-0.00476	0.00994	0.019293	0.054871

Valores de optimización:

SSE	Correlación	Mixing
588.74504	0.99318553	0.00168695

Grafica del Modelo:

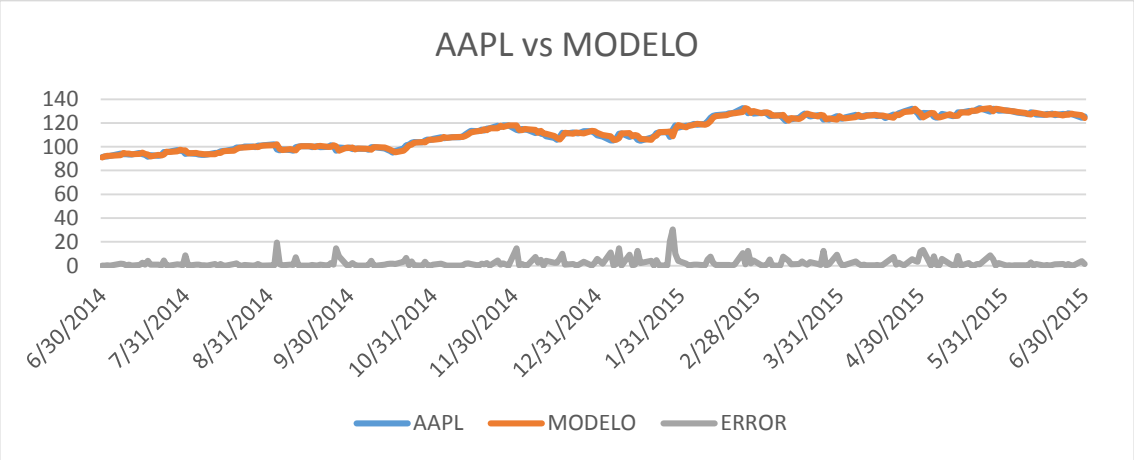

Figure 4-3 Modelo al Optimizar el SSE

4.2.4.2 Ecuación del modelo optimizando la correlación

Valores de coeficientes:

A1	A2	A3	A4	A5	A6	A7
3.00946	-0.10298407	0.22231538	-0.03114	0.111694	0.019237	0.044574

Valores de optimización:

Podemos notar que al momento de realizar la optimización el valor de la suma de errores al cuadrado (SSE) se incrementa bastante, por lo que este modelo queda descartado de la siguiente fase de evaluación.

SSE	Correlación	Mixing
12359553	0.99328869	8.0366E-08

Grafica del Modelo:

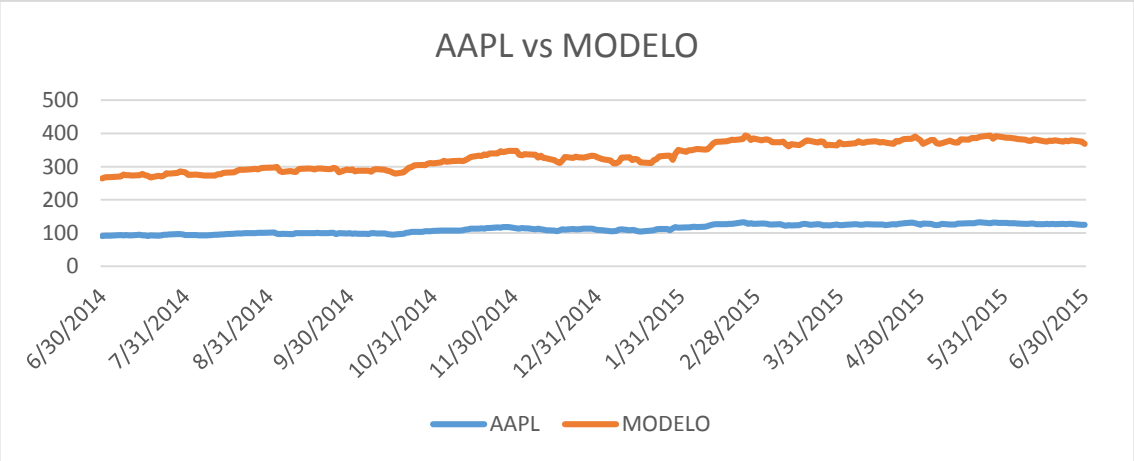

Figure 4-4 Modelo al optimizar la correlación

4.2.4.3 Ecuación del modelo optimizando el mixing

Valores de coeficientes:

A1	A2	A3	A4	A5	A6	A7
0.9710911	-0.05413075	0.04593286	0.00874	0.01123	0.020469	0.05866

Valores de optimización:

SSE	Correlación	Mixing
588.64571	0.99318685	0.00168724

Grafica del Modelo:

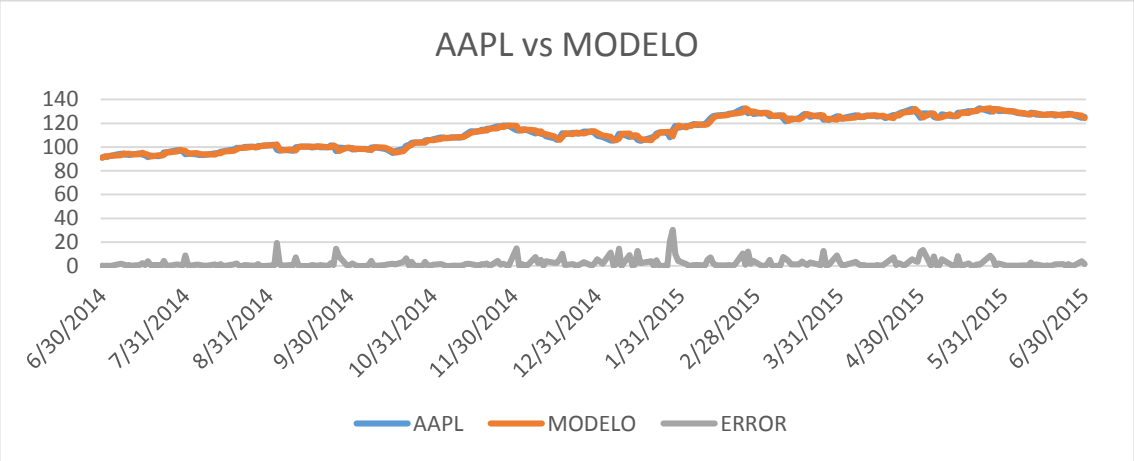

Figure 4-5 Modelo al optimizar el mixing

4.3 Fase V: Evaluación

4.3.1 Evaluar el resultado de *data mining* respecto a los criterios de éxito del negocio

Usaremos los siguientes modelos:

- Modelo sin optimizar

$$APL = 0.987545 * AAPL\_1 - 0.0391297 * CTSH\_3 - 0.0848414 * INTU\_3 + 0.101999 * INTU\_4 + 0.0708764 * SWKS\_1 - 0.0586335 * SWKS\_5 + 0.0188428 * CHKP\_4$$

- Modelo optimizado al minimizar el SSE

$$APL = 0.9717144 * AAPL_1 + 0.0537326 * CTSH_3 - 0.02967956 * INTU_3 - 0.00476 * INTU\_4 + 0.00994 * SWKS\_1 - 0.019293 * SWKS\_5 + 0.054871 * CHKP\_4$$

- Modelo optimizando al maximizar el mixing

$$APL = 0.9710911 * AAPL_1 + 0.05413075 * CTSH\_3 - 0.04593286 * INTU\_3 + 0.00874 * INTU\_4 + 0.01123 * SWKS\_1 - 0.020469 * SWKS\_5 + 0.05866 * CHKP\_4$$

Estos modelos se probaran usando dos estrategias de compraventa de acciones, una denominada toro y otra oso(Sonono & Mashele, 2013).

- Toro es una estrategia alcista formada por lo compra cuando las acciones bajan con la esperanza que el mercado vuelva a subir.

Las condiciones que usaremos para para probar esta estrategia son las siguientes:

- Sí el valor de hoy es menor al valor de mañana vendo.

$$If\left(AAPL_{\_1} < AAPL_{modelo}\right) Then\ VENDER$$

- Sí el valor de hoy es mayor al de mañana compro.

$$If\left(AAPL_{\_1} > AAPL_{modelo}\right) Then\ COMPRAR$$

- Oso es una estrategia bajista formada por la compra cuando las acciones suben de valor

Las condiciones que usaremos para para probar esta estrategia son las siguientes:

- Sí el valor de hoy es menor al valor de mañana compro.

$$If\left(AAPL_{\_1} < AAPL_{modelo}\right) Then\ COMPRAR$$

- Sí el valor de hoy es mayor al de mañana vendo.

$$If\left(AAPL_{\_1} > AAPL_{modelo}\right) Then\ VENDER$$

Para ser probadas con éxito es necesario generar el ambiente de prueba, para esto se propone lo siguiente:

El usuario inicia con un portafolio con 100 acciones de Apple las cuales pueden crecer o disminuir a ritmo de +/- acciones por movimiento y 10,000.00 USD para invertir en la compra de más acciones de la misma compañía en caso que el modelo así lo recomiende o crecer en efectivo como cuando el modelo recomiende vender las acciones.

Nuestro periodo de evaluación es de 1 año de duración (30 de Junio del 2014 al 30 de Junio del 2015)

Por lo que las condiciones iniciales el día cierre del día 29 de junio del 2014 son las siguientes. Tabla 4-2:

- Valor de acción de Apple (AAPL) : 90.393781
- Número de acciones de Apple (AAPL) en el portafolio : 100
- Valor en dólares de las acciones : $ 9,039.3781 USD
- USD disponible para invertir: $10,000.00 USD
- Balance total= Valor de las acciones + USD disponibles para invertir= $19,039.3781

Tabla 4-2 Condiciones iniciales de la evaluación

AAPL	MODELO	ESTRATEGIA	N° ACCIONES	A_VENTA	A_COMPRA	VALOR_$_A	CUENTA_$	BALANCE_$	CAMBIO_%
			100			9039.3781	$ 10,000.00	$19,039.38	0

Para poder llevar a cabo la evaluación diseñamos las siguientes columnas de condiciones de la prueba con base a la recomendación del modelo:

- N° ACCIONES: Columna que muestra el número de acciones disponibles en el portafolio, si la estrategia propuesta por el modelo es VENDER es necesario saber si el portafolio cuenta con los títulos necesarios y en caso contrario sumar los títulos comprados.

$$if((Estrategía = \text{"VENDER"}) \text{ AND } (\text{N° ACCIONES} > 0))$$
$$THEN\ (N°\ ACCIONES - A_{VENTA})$$
$$ELSE\ (N°\ ACCIONES + A_{COMPRA})$$
$$end\ if$$

- A_VENTA: Columna que muestra el número de acciones a vender (cada movimiento involucra la cantidad de 10 títulos), para llevar a cabo la operación es necesario conocer si disponemos de suficientes acciones en nuestro portafolio.

$$if((Estrategía = \text{"VENDER"}) \text{ AND } (\text{N° ACCIONES} > 0))$$
$$THEN\ (10)$$
$$ELSE\ (0)$$
$$end\ if$$

- A_COMPRA: Columna que muestra el número de acciones a comprar (cada movimiento involucra la cantidad de 10 títulos), para llevar a cabo la operación es necesario conocer si disponemos del suficiente dinero (USD) en la cuenta.

$$if((Estrategía = \text{"COMPRAR"}) \text{ AND } (\text{CUENTA\_\$} > 10*\text{AAPL}))$$
$$THEN\ (10)$$
$$ELSE\ (0)$$
$$end\ if$$

- VALOR_$_A: Columna que muestra el valor que tiene el total de acciones en el portafolio al cierre actualizado por día de operaciones.

$$N°\ ACCIONES * Valor\ AAPL\ del\ día$$

- CUENTA_$: Columna que muestra la cantidad de efectivo (USD) disponible para comprar acciones o en caso contrario sumar el valor de las acciones vendidas.

$$if(Estrategía = \text{"COMPRAR"})$$
$$THEN\ (CUENTA_{\$} - (A_{COMPRA} * AAPL))$$
$$ELSE\ (CUENTA_{\$} + (A_{VENTA} * AAPL))$$
$$end\ if$$

- CAMBIO: Columna que muestra el porcentaje de cambio del valor total del portafolio respecto al movimiento realizado en base a la estrategia.

$$\frac{balance\ del\ valor\ del\ portafolio\ antes\ del\ movimiento}{balance\ del\ valor\ del\ portafolio\ \ después\ del\ movimiento} - 1$$

Los resultados de la evaluación son los siguientes respecto a la *buy and hold* comprar y retener hasta el final del ciclo el día 30 de junio del 2015, el cual el valor de acción de Apple (AAPL) es de $125.43 USD

Estrategia *buy and hold*

- Número de acciones de Apple (AAPL) en el portafolio : 100
- Valor en dólares de las acciones : $ 12,543.00 USD
- USD cuenta: $10,000.00 USD
- Balance total= Valor de las acciones + USD cuenta= $22,543.00

Modelo sin optimizar

- Estrategia OSO
    - Número de acciones de Apple (AAPL) en el portafolio : 190
    - Valor en dólares de las acciones : $ 23,831.70 USD
    - USD cuenta: $1,926.97 USD
    - Balance total= Valor de las acciones + USD cuenta= $25,758.67
    - Ganancia neta del periodo respecto a la estrategia *buy and hold* = $3,215.67 o 14.265%

- Estrategia TORO
    - Número de acciones de Apple (AAPL) en el portafolio : 10
    - Valor en dólares de las acciones : $ 1,254.30 USD
    - USD cuenta: $18,073.03 USD
    - Balance total= Valor de las acciones + USD cuenta= $19,327.33
    - Pérdida neta del periodo respecto a la estrategia *buy and hold* = $3,215.67 o - 14.265%

Modelo optimizado al minimizar el SSE

- Estrategia OSO
    - Número de acciones de Apple (AAPL) en el portafolio : 0
    - Valor en dólares de las acciones : $ 0.00 USD
    - USD cuenta: $26,252.12 USD
    - Balance total= Valor de las acciones + USD cuenta= $26,252.12
    - Ganancia neta del periodo respecto a la estrategia *buy and hold* = $3,709.12 o 16.454%
- Estrategia TORO
    - Número de acciones de Apple (AAPL) en el portafolio : 150
    - Valor en dólares de las acciones : $ 18,814.50 USD
    - USD cuenta: $120.18 USD
    - Balance total= Valor de las acciones + USD cuenta= $18,934.68
    - Pérdida neta del periodo respecto a la estrategia *buy and hold* = $3,608.32 o -16.006%

Modelo optimizando al maximizar el mixing

- Estrategia OSO
    - Número de acciones de Apple (AAPL) en el portafolio : 0
    - Valor en dólares de las acciones : $ 0.00 USD
    - USD cuenta: $26,140.95 USD
    - Balance total= Valor de las acciones + USD cuenta= $26,140.95
    - Ganancia neta del periodo respecto a la estrategia *buy and hold* = $3,597.95 o 15.960%

- Estrategia TORO
    - Número de acciones de Apple (AAPL) en el portafolio : 150
    - Valor en dólares de las acciones : $ 18,814.50 USD
    - USD cuenta: $231.35 USD
    - Balance total= Valor de las acciones + USD cuenta= $19,045.85
    - Pérdida neta del periodo respecto a la estrategia *buy and hold* = $(3,497.15) o -15.513%

Tabla 4-3 Evaluación final de las estrategias VS buy and hold

Modelo_Estrategia	Balance total USD	Ganancia	Porcentaje
Buy and hold	$22,543.00	0	0%
Sin optimizar_OSO	$25,758.67	$3,215.67	14.265%
Sin optimizar_TORO	$19,327.33	-$3,215.67	-14.265%
Optimizar SEE_OSO	$26,252.12	$3,709.12	16.454%
Optimizar SEE_TORO	$18,934.68	-$3,608.32	-16.006%
Optimizar Mixing_OSO	$26,140.95	$3,597.95	15.960%
Optimizar Mixing_TORO	$19,045.85	-$3,497.15	-15.513%

4.4 Fase VI: Implementación

4.4.1 Plan de implementación

Esta es la última fase de la metodología de CRISP-DM, que consiste en la implementación, mantenimiento y actualización del modelo, así como generar reportes y documentación para el usuario. Este trabajo de fin de máster no contempla la implementación del modelo generado por la investigación, pero se realizan las siguientes recomendaciones:

La manera de implementar el sistema es usando las condiciones propuestas, aunado al modelo, realizar la programación de una aplicación de un macro en Microsoft Excel[14] de manera que pueda estar actualizando los datos del modelo de manera diaria.

[14] https://products.office.com/en-us/excel

5 Capítulo: Discusión

5.1 Discusión

Con el desarrollo del modelo presentado en este trabajo de fin de máster, se buscaba definir un modelo genérico para realizar una aproximación valida y aplicable del valor de las acciones de una compañía en la bolsa de valores. El modelo propuesto fue evaluado siguiendo dos estrategias diferentes de compraventa de acciones Oso y Toro(Mehmood & Hanif, 2014).Ambas estrategias presentan, de forma simultanea diferencias considerables que influyen de forma decisiva en la aplicación del método y en los resultados obtenidos.

La identificación y extracción de información necesaria para crear la base de datos, fue necesario clasificar una lista de categorías y dominios de la empresas en el caso del presente trabajo fue el NASDAQ en la categoría de Tecnología, además la información se extrajo de forma directa del sitio oficial[15].Esto implica que, para detectar un recurso y la información asociada al mismo, de forma obligada tiene que haber sido publicado e indexado en la web. De otra forma, el recurso nunca será identificado. Esta limitación compensa debido a que garantiza la calidad y veracidad de la información tratada y extraída. Eso lo comprobamos en la sección referente a la calidad de datos (3.3.3), en donde no se tuvo ninguna falta o error en los datos. Esto influyo de manera positiva durante el desarrollo de modelo de predicción del valor de las acciones.

Otro apartado fundamental fue considerar las teorías económicas modernas como el EMH(Fama, 1970), Teoría del andar aleatorio(B. G. Malkiel, 1973), así como de otros tantos, es plausible que cada autor tengo su propia teoría la cual tiene una limitación de aplicación por lo que es perfectamente válido realizar una hipótesis y poder validarla a través del desarrollo de este trabajo de investigación con sus propias limitaciones.

[15] https://data.nasdaq.com/

Al analizar los resultados de la aplicación de la estrategia al modelo construido mediante regresión y optimizado posteriormente.se ha observado una diferencia significativa entre ambos sistemas. Los datos parecen indicar que la estrategia bajista y conservadora conocida en el trabajo como Oso obtiene una ventaja significativa al momento de conservar el valor total del portafolio.

5.2 Otras cuestiones metodológicas

El modelo propuesto obliga a plantearse una serie de cuestiones fundamentales antes de comenzar, que repercuten de manera decisiva, en los objetivos. La primera cuestión es la selección de datos. Como ya se mencionó en las primeras secciones de este trabajo, esta tarea es crucial y los resultados obtenidos dependerán de la misma. La segunda cuestión es el alcance y generalidad que se quiera lograr del modelo obtenido, es este caso fue usando sólo una compañía como objetivo en un portafolio con una única variable de acción, toda la información cubriendo siempre una única industria dentro de una sola bolsa de valores, algunos trabajos buscan generar modelos que puedan predecir el valor de un portafolio de múltiples acciones (Moon & Yao, 2011) que parecen tener buenos resultados en el ambiente académico, pero para el usuario que no es experto en la materia le resulta prácticamente imposible aplicarlo en su entorno.

El sistema de predicción basado en datos históricos, se basa en la premisa de que se puede repetir la historia y el comportamiento de una acción en base a este y a otras compañías del entorno, se intenta simplificar la manera de generar el modelo, así como la elección de los datos necesarios para llevar a cabo el experimento, sin perder de vista la precisión y fiabilidad que es lo que importa al inversionista.

Siguiendo con esto, cabe destacar que el modelo propuesto, en el presente trabajo de fin de máster, puede automatizar parcialmente el proceso de seguimiento de una acción en la bolsa de valores. Sim embargo, en la actualidad, en las implementaciones que se llevan a cabo(Bosire, 2014) todavía existen ciertas tareas que son realizadas de forma manual, por lo que no se podría hablar, en estos momentos, de realizar un método completamente automático.

Algunas de estas tareas son por ejemplo, la revisión de la información extraída y la clasificación de la misma de suma importancia para construir el modelo.

Finalmente, comentar que, este modelo no aborda, de manera explícita, el hecho de la actualización de los contenidos necesarios para la creación del modelo, ya que al ser una bolsa de valores es posible que algunas compañías dejen de operar en un futuro, para estos casos sólo será necesario repetir el proceso de actualización de la base de datos para aquellos cambios en las compañías a lo largo del tiempo, esto basta con consultar de forma periódica las fuentes de información seleccionada para este fin.

6 Capítulo: Conclusiones y líneas futuras

6.1 Conclusiones finales

Como objetivo principal de este trabajo de fin de máster se planteaba la demostración de la hipótesis planteada en la sección 1.3.Esta hipótesis denotaba lo siguiente:

“Realizando un análisis técnico de los datos históricos del valor de las acciones durante un plazo determinado, es posible predecir el cambio de precio de una acción, con la finalidad de ayudar al inversor al momento de tomar la decisión de compraventa *trading* de sus acciones, utilizando para esto regresión múltiple”

Para la verificación empírica de esta hipótesis, se ha definido un modelo que ha sido evaluado usando dos estrategias las cuales de acuerdo a la tabla de evaluación 4-3.permite demostrar de manera exitosa que el modelo optimizado usando una estrategia denominada bajista u oso permiten obtener un margen bastante significativo de ganancia respecto a sólo usar la estrategia de comprar y mantener la acción durante un periodo de tiempo determinado, este valor de ganancia tiene un rango entre 14.265% para el modelo sin optimizar y 16.454% para el modelo optimizado de forma que minimice el SSE.

La forma de construir el modelo, permite reutilizarlo para realizar predicciones de otras compañías en la misma bolsa de valores, sólo teniendo cuidado de elegir un dominio de acciones de la misma industria para evitar cambios bruscos o no esperados en los resultados.

6.2 Líneas futuras de investigación

Para terminar con este último capítulo del trabajo de fin de máster, en este apartado se enumeran una serie de líneas futuras de investigación que se desprenden del trabajo de investigación presentado.

En general, las líneas futuras se basan en la ampliación de algunas características concentras del modelo desarrollado y la posible automatización

del proceso, que ahora se realiza de manera manual. Las potenciales líneas de investigación son las siguientes:

- Realizar experimento y prueba con acciones de otras compañías. El método utilizado para obtener el modelo de estimación del valor de la acción de la compañía Apple (AAPL) es genérico de manera que puede ser usado para cualquier compañía dentro del NASDAQ, la complejidad de esto sólo radica en obtener la información necesaria para replicar el experimento.
- Automatizar la tarea de actualización de datos en caso que algún atributo para generar el modelo no se encuentre disponible, de manera que se pueda continuar con el *trading* sin miedo a perder de vista las futuras recomendaciones realizadas por el modelo.
- Crear un modelo genérico para poder llevar a cabo un experimento en un portafolio de múltiples compañías en diferentes bolsas de valores alrededor del mundo.
- Extender la manera de obtención de datos usando también los datos fundamentales, como lo son reportes internos de las compañías y el uso de text mining en la web, como puede ser *blogs,* páginas web, redes sociales, etc. Esto con la intención de identificar noticias relacionadas a la compañía que pudiera afectar el desempeño del valor la acción, esto en el menor lapso de tiempo posible para ganar ventaja frente a otros inversionistas.
- Realizar pruebas con diferentes perfiles de inversión de manera que se obtenga estrategias personalizas para el inversionista.

Anexo a

Modelos

Mínimos cuadrados ordinarios con constante

R-squared = 99.7827 percent
R-squared (adjusted for d.f.) = 99.752 percent
Standard Error of Est. = 0.989714
Mean absolute error = 0.660685
Durbin-Watson statistic = 2.00549 (P=0.4613)
Lag 1 residual autocorrelation = -0.00296146

La ecuación del modelo es:

AAPL = 0.777211 + 0.949318*AAPL_1 - 0.0317236*AAPL_2 - 0.0235243*AAPL_3 + 0.103096*AAPL_4 - 0.0331009*AAPL_5 - 0.00816527*GOOGL_1 + 0.00658365*GOOGL_2 + 0.00318112*GOOGL_3 - 0.000432833*GOOGL_4 - 0.00378394*GOOGL_5 - 0.0251846*MSFT_1 + 0.00231233*MSFT_2 + 0.163018*MSFT_3 - 0.103062*MSFT_4 + 0.0859578*MSFT_5 + 0.1663*INTC_1 - 0.364686*INTC_2 + 0.216947*INTC_3 - 0.138741*INTC_4 + 0.0330905*INTC_5 + 0.0618707*CSCO_1 + 0.113205*CSCO_2 - 0.0953656*CSCO_3 - 0.102579*CSCO_4 + 0.0328207*CSCO_5 + 0.056584*QCOM_1 - 0.185733*QCOM_2 + 0.140592*QCOM_3 - 0.094571*QCOM_4 + 0.0888797*QCOM_5 - 0.096554*TXN_1 + 0.150609*TXN_2 - 0.0126784*TXN_3 - 0.190243*TXN_4 + 0.104837*TXN_5 + 0.0025826*BIDU_1 + 0.00855276*BIDU_2 + 0.00182039*BIDU_3 - 0.0213011*BIDU_4 + 0.0120134*BIDU_5 + 0.0871179*ADBE_1 - 0.103392*ADBE_2 + 0.0715566*ADBE_3 - 0.0590237*ADBE_4 + 0.00406988*ADBE_5 - 0.0662833*ADP_1 + 0.160934*ADP_2 + 0.000746214*ADP_3 + 0.0252674*ADP_4 - 0.161734*ADP_5 + 0.105127*CTSH_1 - 0.194654*CTSH_2 + 0.00248834*CTSH_3 + 0.00286936*CTSH_4 + 0.00830171*CTSH_5 + 0.0481584*YHOO_1 - 0.0401168*YHOO_2 + 0.029444*YHOO_3 - 0.143628*YHOO_4 + 0.0957514*YHOO_5 + 0.002184*BRCM_1 + 0.0223262*BRCM_2 - 0.00881141*BRCM_3 - 0.047343*BRCM_4 - 0.00793906*BRCM_5 - 0.0469203*INTU_1 + 0.101864*INTU_2 - 0.173433*INTU_3 + 0.163854*INTU_4 - 0.0076272*INTU_5 - 0.0258384*CERN_1 + 0.0122528*CERN_2 - 0.0572541*CERN_3 + 0.0570669*CERN_4 - 0.0118488*CERN_5 - 0.0959009*AMAT_1 + 0.341279*AMAT_2 - 0.216184*AMAT_3 - 0.210131*AMAT_4 + 0.215254*AMAT_5 - 0.0394857*EA_1 + 0.0811881*EA_2 - 0.00696058*EA_3 - 0.044022*EA_4 - 0.0724938*EA_5 + 0.00619842*FISV_1 - 0.0299541*FISV_2 - 0.0606333*FISV_3 - 0.00279643*FISV_4 + 0.144262*FISV_5 - 0.0554161*ADI_1 - 0.0289452*ADI_2 - 0.213232*ADI_3 + 0.276685*ADI_4 + 0.0106883*ADI_5 - 0.110126*MU_1 + 0.20072*MU_2 - 0.193899*MU_3 + 0.238659*MU_4 - 0.139847*MU_5 + 0.0429105*SWKS_1 - 0.0267488*SWKS_2 + 0.165407*SWKS_3 - 0.0867233*SWKS_4 - 0.0836804*SWKS_5 + 0.0145173*WDC_1 - 0.0291834*WDC_2 + 0.0445285*WDC_3 + 0.121464*WDC_4 - 0.160872*WDC_5 + 0.300812*ATVI_1 - 0.341249*ATVI_2 + 0.328994*ATVI_3 + 0.0203607*ATVI_4 - 0.233211*ATVI_5 - 0.16864*SYMC_1 + 0.010767*SYMC_2 + 0.19544*SYMC_3 + 0.117076*SYMC_4 - 0.164623*SYMC_5 + 0.138737*ALTR_1 - 0.12995*ALTR_2 - 0.0105988*ALTR_3 + 0.062226*ALTR_4 - 0.0114865*ALTR_5 + 0.0885617*STX_1 - 0.154698*STX_2 + 0.0328*STX_3 - 0.0674101*STX_4 + 0.102267*STX_5 + 0.00566129*CHKP_1 + 0.0250395*CHKP_2 - 0.0979596*CHKP_3 + 0.117392*CHKP_4 - 0.00196554*CHKP_5 - 0.026444*CA_1 + 0.266372*CA_2 + 0.0775833*CA_3 - 0.286799*CA_4 + 0.00843924*CA_5 - 0.0251153*LRCX_1 + 0.00822053*LRCX_2 - 0.0198026*LRCX_3 + 0.0245976*LRCX_4 - 0.0116474*LRCX_5 - 0.0337288*ADSK_1 + 0.122856*ADSK_2 - 0.0755757*ADSK_3 - 0.0158264*ADSK_4 - 0.0107031*ADSK_5 - 0.0226504*SNDK_1 + 0.0288812*SNDK_2 - 0.0196035*SNDK_3 + 0.0302286*SNDK_4 + 0.0151854*SNDK_5

Mínimos cuadrados ordinarios sin constante

R-squared = 99.9753 percent
R-squared (adjusted for d.f.) = 99.9718 percent
Standard Error of Est. = 0.989472
Mean absolute error = 0.660397
Durbin-Watson statistic = 2.00626
Lag 1 residual autocorrelation = -0.00331749

La ecuación del modelo es:

AAPL = 0.949232*AAPL_1 - 0.0325928*AAPL_2 - 0.0231328*AAPL_3 + 0.102976*AAPL_4 - 0.0334706*AAPL_5 - 0.00819229*GOOGL_1 + 0.00670444*GOOGL_2 + 0.00307144*GOOGL_3 - 0.000322568*GOOGL_4 - 0.0041199*GOOGL_5 - 0.0227151*MSFT_1 + 0.00415092*MSFT_2 + 0.161262*MSFT_3 - 0.103823*MSFT_4 + 0.0894896*MSFT_5 + 0.168786*INTC_1 - 0.364616*INTC_2 + 0.218689*INTC_3 - 0.139245*INTC_4 + 0.0477217*INTC_5 + 0.0640197*CSCO_1 + 0.114047*CSCO_2 - 0.0987302*CSCO_3 - 0.10225*CSCO_4 + 0.0368574*CSCO_5 + 0.0565838*QCOM_1 - 0.185532*QCOM_2 + 0.142165*QCOM_3 - 0.0945977*QCOM_4 + 0.0898107*QCOM_5 - 0.103429*TXN_1 + 0.147297*TXN_2 - 0.0141974*TXN_3 - 0.188931*TXN_4 + 0.101295*TXN_5 + 0.00234856*BIDU_1 + 0.00843374*BIDU_2 + 0.00179639*BIDU_3 - 0.0213596*BIDU_4 + 0.0128986*BIDU_5 + 0.0893779*ADBE_1 - 0.104066*ADBE_2 + 0.0722736*ADBE_3 - 0.0585453*ADBE_4 + 0.00473802*ADBE_5 - 0.0670526*ADP_1 + 0.160929*ADP_2 + 0.00201108*ADP_3 + 0.0257316*ADP_4 - 0.16167*ADP_5 + 0.109071*CTSH_1 - 0.19351*CTSH_2 + 0.00430634*CTSH_3 + 0.00305862*CTSH_4 + 0.00807662*CTSH_5 + 0.050565*YHOO_1 - 0.0393213*YHOO_2 +0.0303817*YHOO_3 - 0.143047*YHOO_4 + 0.0966695*YHOO_5 + 0.00629412*BRCM_1 + 0.0225456*BRCM_2 - 0.00794183*BRCM_3 - 0.0475445*BRCM_4 - 0.0066899*BRCM_5 - 0.044399*INTU_1 + 0.102074*INTU_2 - 0.173376*INTU_3 + 0.164816*INTU_4 - 0.00757431*INTU_5 - 0.0266611*CERN_1 + 0.0121404*CERN_2 - 0.0573839*CERN_3 + 0.0573049*CERN_4 - 0.0134162*CERN_5 - 0.100535*AMAT_1 + 0.343698*AMAT_2 - 0.21612*AMAT_3 - 0.209523*AMAT_4 + 0.224382*AMAT_5 - 0.0378105*EA_1 + 0.081531*EA_2 - 0.00658465*EA_3 - 0.0434152*EA_4 - 0.0727614*EA_5 + 0.00837315*FISV_1 - 0.0304207*FISV_2 - 0.0616693*FISV_3 - 0.00324413*FISV_4 + 0.149293*FISV_5 - 0.0516591*ADI_1 - 0.0274616*ADI_2 - 0.212525*ADI_3 + 0.275076*ADI_4 + 0.00724713*ADI_5 - 0.113121*MU_1 + 0.2004*MU_2 - 0.193775*MU_3 + 0.237111*MU_4 - 0.141813*MU_5 + 0.0436652*SWKS_1 - 0.0272555*SWKS_2 + 0.164495*SWKS_3 - 0.0867705*SWKS_4 - 0.0852514*SWKS_5 + 0.0178045*WDC_1 - 0.0298248*WDC_2 + 0.0446514*WDC_3 + 0.122246*WDC_4 - 0.161985*WDC_5 + 0.304247*ATVI_1 - 0.340818*ATVI_2 + 0.326782*ATVI_3 + 0.0170354*ATVI_4 - 0.232467*ATVI_5 - 0.168563*SYMC_1 + 0.010246*SYMC_2 + 0.197482*SYMC_3 + 0.119126*SYMC_4 - 0.165442*SYMC_5 + 0.143377*ALTR_1 - 0.129195*ALTR_2 - 0.0121934*ALTR_3 + 0.0629391*ALTR_4 - 0.0138786*ALTR_5 + 0.0796744*STX_1 - 0.155045*STX_2 + 0.032951*STX_3 - 0.067514*STX_4 + 0.103036*STX_5 + 0.00425228*CHKP_1 + 0.0255398*CHKP_2 - 0.0988313*CHKP_3 + 0.117524*CHKP_4 - 0.00459797*CHKP_5 - 0.035256*CA_1 + 0.26627*CA_2 + 0.0787654*CA_3 - 0.287402*CA_4 + 0.0150444*CA_5 - 0.0247999*LRCX_1 +0.00776653*LRCX_2 - 0.0193921*LRCX_3 + 0.0247305*LRCX_4 - 0.0112731*LRCX_5 - 0.0377044*ADSK_1 + 0.122074*ADSK_2 - 0.0767884*ADSK_3 - 0.0168913*ADSK_4 - 0.00906093*ADSK_5 - 0.0240096*SNDK_1 + 0.0296434*SNDK_2 – 0.0205078*SNDK_3 + 0.0300607*SNDK_4 + 0.0138923*SNDK_5

Selección de paso hacia adelante con constante

R-squared = 99.7461 percent
R-squared (adjusted for d.f.) = 99.7457 percent
Standard Error of Est. = 1.00236
Mean absolute error = 0.681536
Durbin-Watson statistic = 1.9645 (P=0.2649)
Lag 1 residual autocorrelation = 0.0176517

La ecuación del modelo es:

AAPL = -0.572336 + 0.991379*AAPL_1 + 0.0544873*INTC_1

Selección de paso hacia adelante sin constante

R-squared = 99.9715 percent
R-squared (adjusted for d.f.) = 99.9714 percent
Standard Error of Est. = 0.996135
Mean absolute error = 0.676402
Durbin-Watson statistic = 1.9869
Lag 1 residual autocorrelation = 0.00629484

La ecuación del modelo es:

AAPL = 0.987545*AAPL_1 - 0.0391297*CTSH_3 - 0.0848414*INTU_3 + 0.101999*INTU_4 + 0.0708764*SWKS_1 - 0.0586335*SWKS_5 + 0.0188428*CHKP_4

Selección de paso hacia atrás con constante

R-squared = 99.7823 percent
R-squared (adjusted for d.f.) = 99.7624 percent
Standard Error of Est. = 0.968855
Mean absolute error = 0.661034
Durbin-Watson statistic = 2.00465 (P=0.4672)
Lag 1 residual autocorrelation = -0.00252882

La ecuación del modelo es:

AAPL = 0.886926 + 0.946811*AAPL_1 - 0.0269961*AAPL_2 - 0.0279695*AAPL_3 + 0.105948*AAPL_4 - 0.0337822*AAPL_5 - 0.00706709*GOOGL_1 + 0.00452234*GOOGL_2 + 0.00420607*GOOGL_3 - 0.00412985*GOOGL_5 + 0.14644*MSFT_3 - 0.101677*MSFT_4 + 0.082242*MSFT_5 + 0.161242*INTC_1 - 0.373984*INTC_2 + 0.233046*INTC_3 - 0.114926*INTC_4 + 0.0577772*CSCO_1 + 0.120019*CSCO_2 - 0.0976611*CSCO_3 - 0.069821*CSCO_4 + 0.0568366*QCOM_1 - 0.190524*QCOM_2 + 0.146481*QCOM_3 - 0.0949496*QCOM_4 + 0.0876935*QCOM_5 - 0.0963256*TXN_1 + 0.137212*TXN_2 - 0.18647*TXN_4 + 0.100576*TXN_5 + 0.0115415*BIDU_2 - 0.018983*BIDU_4 + 0.0109964*BIDU_5 + 0.0823835*ADBE_1 - 0.104175*ADBE_2 + 0.0764878*ADBE_3 - 0.056047*ADBE_4 - 0.0730974*ADP_1 + 0.166255*ADP_2 - 0.135764*ADP_5 + 0.109873*CTSH_1 - 0.19155*CTSH_2 + 0.0315874*YHOO_1 - 0.143997*YHOO_4 + 0.100618*YHOO_5 - 0.0420467*BRCM_4 - 0.0532817*INTU_1 + 0.105336*INTU_2 - 0.174062*INTU_3 + 0.161018*INTU_4 - 0.0232105*CERN_1 - 0.0515628*CERN_3 + 0.048222*CERN_4 - 0.0866735*AMAT_1 + 0.339231*AMAT_2 - 0.25748*AMAT_3 - 0.163362*AMAT_4 + 0.206489*AMAT_5 - 0.0408635*EA_1 + 0.0696086*EA_2 - 0.0395549*EA_4 - 0.0740241*EA_5 - 0.0754359*FISV_3 + 0.135941*FISV_5 - 0.0614709*ADI_1 - 0.241212*ADI_3 + 0.291336*ADI_4 - 0.118697*MU_1 + 0.211889*MU_2 - 0.199067*MU_3 + 0.230883*MU_4 - 0.125272*MU_5 + 0.0326821*SWKS_1 + 0.151437*SWKS_3 - 0.0904801*SWKS_4 - 0.084017*SWKS_5 - 0.0258118*WDC_2 + 0.0590602*WDC_3 + 0.111397*WDC_4 - 0.156225*WDC_5 + 0.297476*ATVI_1 - 0.337736*ATVI_2 + 0.350075*ATVI_3 - 0.236687*ATVI_5 - 0.163384*SYMC_1 + 0.210894*SYMC_3 + 0.109456*SYMC_4 - 0.166096*SYMC_5 + 0.137065*ALTR_1 - 0.131205*ALTR_2 + 0.0440582*ALTR_4 + 0.0962862*STX_1 - 0.142054*STX_2 - 0.0574526*STX_4 + 0.1055*STX_5 + 0.0256663*CHKP_2 - 0.0907535*CHKP_3 + 0.116075*CHKP_4 + 0.278622*CA_2 - 0.240344*CA_4 - 0.0229976*LRCX_1 - 0.0311905*ADSK_1 + 0.114694*ADSK_2 - 0.0701727*ADSK_3 - 0.0229202*ADSK_4 + 0.0329694*SNDK_4

Selección de paso hacia atrás sin constante

R-squared = 99.9752 percent
R-squared (adjusted for d.f.) = 99.973 percent
Standard Error of Est. = 0.968724
Mean absolute error = 0.661022
Durbin-Watson statistic = 2.00528
Lag 1 residual autocorrelation = -0.00284394

La ecuación del modelo es:

AAPL = 0.946637*AAPL_1 - 0.028695*AAPL_2 - 0.0263986*AAPL_3 + 0.105637*AAPL_4 - 0.0338931*AAPL_5 - 0.00720108*GOOGL_1 + 0.0048762*GOOGL_2 + 0.00412868*GOOGL_3 - 0.00461389*GOOGL_5 + 0.153853*MSFT_3 - 0.113443*MSFT_4 + 0.0935727*MSFT_5 + 0.158527*INTC_1 - 0.369057*INTC_2 + 0.229558*INTC_3 - 0.0943086*INTC_4 + 0.0612472*CSCO_1 + 0.113559*CSCO_2 - 0.0989911*CSCO_3 - 0.0617742*CSCO_4 + 0.0562942*QCOM_1 - 0.19254*QCOM_2 + 0.149587*QCOM_3 - 0.0954462*QCOM_4 + 0.0890319*QCOM_5 - 0.107759*TXN_1 + 0.135824*TXN_2 - 0.185969*TXN_4 + 0.0965389*TXN_5 + 0.0110274*BIDU_2 - 0.0188214*BIDU_4 + 0.0116048*BIDU_5 + 0.0823045*ADBE_1 - 0.100973*ADBE_2 + 0.0811218*ADBE_3 - 0.0619851*ADBE_4 - 0.0752878*ADP_1 + 0.172133*ADP_2 - 0.138277*ADP_5 + 0.114202*CTSH_1 - 0.187069*CTSH_2 + 0.035923*YHOO_1 - 0.142498*YHOO_4 + 0.101277*YHOO_5 + 0.0200056*BRCM_2 - 0.054205*BRCM_4 - 0.0480291*INTU_1 + 0.105489*INTU_2 - 0.173393*INTU_3 + 0.159896*INTU_4 - 0.023597*CERN_1 - 0.050431*CERN_3 + 0.0446865*CERN_4 + 0.25652*AMAT_2 - 0.25319*AMAT_3 - 0.172382*AMAT_4 + 0.224498*AMAT_5 - 0.0376481*EA_1 + 0.0742839*EA_2 - 0.0437747*EA_4 - 0.0738805*EA_5 - 0.0748001*FISV_3 + 0.140719*FISV_5 - 0.0603684*ADI_1 - 0.235282*ADI_3 + 0.287182*ADI_4 - 0.118384*MU_1 + 0.21907*MU_2 - 0.19773*MU_3 + 0.219282*MU_4 - 0.128519*MU_5 + 0.032133*SWKS_1 + 0.150528*SWKS_3 - 0.0907296*SWKS_4 - 0.0861714*SWKS_5 - 0.0222312*WDC_2 + 0.0578944*WDC_3 + 0.113099*WDC_4 - 0.157758*WDC_5 + 0.297352*ATVI_1 - 0.340218*ATVI_2 + 0.348254*ATVI_3 - 0.231136*ATVI_5 - 0.164736*SYMC_1 + 0.215727*SYMC_3 + 0.111683*SYMC_4 - 0.169413*SYMC_5 + 0.140259*ALTR_1 - 0.129117*ALTR_2 + 0.0404541*ALTR_4 + 0.0881092*STX_1 - 0.142319*STX_2 - 0.0584823*STX_4 + 0.106048*STX_5 + 0.0252193*CHKP_2 - 0.090121*CHKP_3 + 0.111994*CHKP_4 + 0.269566*CA_2 - 0.233565*CA_4 - 0.0232703*LRCX_1 - 0.0379168*ADSK_1 + 0.113314*ADSK_2 - 0.08992*ADSK_3 - 0.0101277*SNDK_1 + 0.0387888*SNDK_4

Glosario y acrónimos

NASDAQ	Acrónimo de *National Association of Securities Dealers Automated Quotations.*
CRISP-DM	Acronimo de Cross Industry Standard Process for Data Mining
Trading	consiste en comprar o vender un valor subyacente en un mercado financiero con la intención de obtener un beneficio especulativo.
Variable	Es un conjunto de mediciones de la misma característica en determinados individuos de una población.
Coeficiente	Elemento constante en una multiplicación
Macro	Es un conjunto de instrucciones que se ejecutan de manera secuencial por medio de una orden de ejecución
Portafolio	es una determinada combinación de activos financieros en los cuales se invierte
Blogs	es una página web en la que se publican regularmente artículos cortos con contenido actualizado y novedoso sobre temas específicos o libres.

Bibliografía

Alraddadi, R. (2015). Statistical Analysis of Stock Prices in John Wiley & Sons. *Journal of Emerging Trends in Computing and Information Sciences*, *6*(1), 38–47.

Bachelier, L. (1906). *Louis Bachelier's Theory of Speculation: The Origins of Modern Finance*. Princeton University Press.

Bosire, M. (2014). The Effect of Automation on Stock Market Price Volatility : A Case of Nairobi Securities Exchange. *IOSR Journal of Economics and Finance*, *5*(3), 71–79. doi:2321-5933

Brown, S. H. (2009). Multiple Linear Regression Analysis : A Matrix Approach with MATLAB. *Alabama Journal of Mathematics*, 1–3.

Chong, F., Ling, H., Ng, D., Yat, C., & Muhamad, R. (2014). An Empirical Re-Investigation on the " Buy-and-hold Strategy " in Four Asian Markets : A 20 Years ' Study. *World Applied Sciences Journal 30 (Innovation Challenges in Multidiciplinary Research & Practice)*, *30*, 226–237. doi:10.5829/idosi.wasj.2014.30.icmrp.30

Corporation, I. B. M. (2011). IBM SPSS Advanced Statistics 20.

Dubey, P. (2012). Association Rule Mining on Distributed Data, *3*(1), 1–6.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940. doi:10.1016/j.eswa.2005.06.024

Fama, E. (1970). Efficient Capital Markets:A Review of Theory and empirical Work.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI Magazine*, 37–54. doi:10.1145/240455.240463

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. doi:10.1145/240455.240464

Fisher, I. (1907). The Nature of Capital and Income. *Journal of Political Economy*, *15*(3), 129. doi:10.1086/251299

Foster, I., & Kesselman, C. (2004). The Grid in a nutshell. *Grid Resource Management: State of the Art and Future Trends*, 3–13. doi:10.1007/978-1-4615-0509-9_1

Frawley, W. J., Piatetsky-shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases : An Overview. *AI Magazine*, *13*(3), 57–70. doi:10.1609/aimag.v13i3.1011

Greemblatt, J. (2005). *The Little Book That Still Beats the Market* (Vol. 1). Retrieved from Wiley

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. *Annals of Physics*, *54*, 770. doi:10.5860/CHOICE.49-3305

Hand, D. J., & Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician*, *52*(2), 112. doi:10.2307/2685468

Hansen, B. (1999). Discussion of "Data mining reconsidered." *The Econometrics Journal*, *2*, 192 – 201. doi:10.1111/1368-423X.00026

Hellstrom, T. (1998). *A random Walk through the stock market*.

Ibm. (2010). 8_CRISP-DM 1.0 Step by stop data mining guide. *IBM Corporation*. Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03084usen/YTW03084USEN.PDF

Kock, N., & Verville, J. (2012). Exploring Free Questionnaire Data with Anchor Variables. *International Journal of Healthcare Information Systems and Informatics*, *7*(1), 46–63. doi:10.4018/jhisi.2012010104

Kroha, P., & Baeza-Yates, R. (2004). Classification of stock exchange news. *Technical Report*.

Malkiel, B., & Fama, E. (1970). Efficient capital markets: A review of theory and empirical work*. *The Journal of Finance*, *25*(2), 28–30. doi:10.1111/j.1540-6261.1970.tb00518.x

Malkiel, B. G. (1973). A Random Walk Down Wall Street. *Foundations*. doi:10.1111/1467-6419.00091

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. doi:10.1111/j.1540-6261.1952.tb01525.x

Mehmood, Y., & Hanif, W. (2014). Impact of Bullish and Bearish Market on Investor Sentiment, *9*(1), 142–151.

Moon, Y., & Yao, T. (2011). A robust mean absolute deviation model for portfolio optimization. *Computers & Operations Research*, *38*(9), 1251–1258. doi:10.1016/j.cor.2010.10.020

Murphy, J. J. (1999). *Technical Analysis Of The Financial Markets*.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. *Springer*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-76917-0

Perlich, C., Dalessandro, B., Hook, R., Stitelman, O., Raeder, T., & Provost, F. (2012). Bid optimizing and inventory scoring in targeted online advertising. *Proceedings of the ACM SIGKDD 2012*, 804. doi:10.1145/2339530.2339655

Phillips, P. C. B., Haven, N., & Foundation, C. (2006). in Econometrics By Cowles Foundation for Research in Economics Automated Discovery, (1149).

Schumaker, R. P., & Chen, H. (2008). Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies. *Journal of the American Society for Information Science and Technology*, *59*(2), 247–255. doi:10.1002/asi.20739

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing and Management*, *45*(5), 571–583. doi:10.1016/j.ipm.2009.05.001

Schumaker, R. P., & Chen, H. (2010). Textual Analysis of Stock Market Prediction Using Financial News Articles. *The Technology World …*, 1–20. Retrieved from http://ismeindia.com/Downloads/Papers/THE_TECHNOLOGY_WORLD.pdf#page=36

Seng, D., & Hancock, J. R. (2012). Fundamental Analysis and the Prediction of Earnings. *International Journal of Business and Management*, *7*(3), 32–46. doi:10.5539/ijbm.v7n3p32

Sonono, M., & Mashele, H. (2013). Assessing the Risks of Trading Strategies Using Acceptability Indices. *Journal of Mathematical Finance*, *2013*(November), 465–475. Retrieved from http://www.scirp.org/journal/PaperInformation.aspx?PaperID=40089&

Tables, T. C. (2009). 7_Multiple Regression, 1–20.

Thomsett, M. C. (1998). Mastering Fundamental Analysis. *Economic Indicators*, 242.

Tsang, P. M., Kwok, P., Choy, S. O., Kwan, R., Ng, S. C., Mak, J., … Wong, T.-L. (2007, June). Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence*. doi:10.1016/j.engappai.2006.10.002

WANG, J., & CHAN, S. (2006). Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications*, *30*(4), 605–611. doi:10.1016/j.eswa.2005.07.006

Wang, Y. (2003). Mining stock price using fuzzy rough set system. *Expert Systems with Applications*, *24*(1), 13–23. doi:10.1016/S0957-4174(02)00079-9